An Explainable Framework for Biomedical Text Classification Combining Regular Expressions and Machine Learning Models

Facultad de Ciencias Empresariales
Universidad del Bío-Bío

Mauricio Fuenzalida
mauricio.fuenzalida2401@alumnos.ubiobio.cl

Profesores Guía
Rosa L. Figueroa Christopher A. Flores

Resumen

La clasificación de textos biomédicos es esencial para la organización automática de la información clínica. Sin embargo, su implementación sigue siendo un desafío debido a la transparencia limitada de los clasificadores. Para abordar estos problemas, este estudio propone un sistema híbrido para la clasificación de textos clínicos que integra clasificadores tradicionales y modernos con reglas lógicas interpretables basadas en expresiones regulares. El objetivo principal es abordar la falta de transparencia en los modelos, que a menudo funcionan como "cajas negras" difíciles de interpretar. Con este fin, se diseñó una arquitectura que aprovecha dos fuentes de información complementarias: patrones generados automáticamente derivados de expresiones regulares y representaciones estadísticas y contextuales derivadas de modelos tradicionales y basados en Transformer. Para verificar que la arquitectura aprende conceptos clínicos en lugar ruido estadístico, se aplicaron técnicas de explicabilidad. Además, se realizaron pruebas de enmascaramiento en las que se ocultaron términos clave para medir cuánto disminuía el rendimiento del modelo sin esa información esencial. Los experimentos realizados en tres conjuntos de datos, el Chilean Waiting List Corpus (CWLC), el Medical Information Mart for Intensive Care III (MIMIC-III) y el Medical Abstracts Text Classification Dataset, demuestran que la inclusión de reglas lógicas mejoró el desempeño de los modelos tradicionales hasta en un 4,5 % en términos de la métrica valor-F1. Además, el análisis de explicabilidad reveló que los modelos mantienen un desempeño robusto incluso ante la pérdida de información esencial, donde el enmascaramiento de términos clave provocó caídas de rendimiento de hasta el 50 %, confirmando que el sistema basa sus predicciones en términos clínicamente relevantes en lugar de ruido estadístico. Este trabajo concluye que es posible obtener clasificadores de alto rendimiento que sean técnicamente confiables y puedan ser revisados por expertos.

Palabras clave: clasificación de textos biomédicos, inteligencia artificial explicable, expresiones regulares

Índice

1. Definición del Problema de Estudio y Oportunidad

El procesamiento de grandes volúmenes de datos en registros de salud electrónicos y literatura científica ha impulsado el uso de modelos de aprendizaje profundo, particularmente arquitecturas basadas en Transformers, para la clasificación automática de textos biomédicos. Sin embargo, la implementación práctica de estas tecnologías en entornos clínicos enfrenta una desafío crítico: su naturaleza de "caja negra" [10].

Esta falta de transparencia dificulta que los expertos clínicos y los organismos reguladores comprendan y validen los mecanismos internos que llevan a una decisión específica. En un escenario donde los errores pueden tener consecuencias clínicas graves, la dependencia de modelos puramente estadísticos que pueden capturar sesgos en lugar de patrones lingüísticos reales genera desconfianza y limita la adopción tecnológica. Por otro lado, aunque las expresiones regulares (RegExes) ofrecen una transparencia intrínseca y legible por humanos, su diseño manual es una tarea laboriosa que requiere un conocimiento profundo del dominio y del lenguaje técnico [2].

Surge entonces la oportunidad de desarrollar marcos de trabajo híbridos y explicables que integren la potencia predictiva de los modelos de aprendizaje automático y los Transformers como *Biomedical Bidirectional Encoder Representations from Transformers* (BioBERT) y *Sentence Transformer Fine-Tuning* (SetFit), con la interpretabilidad de reglas lógicas generadas automáticamente.

La automatización en la extracción de patrones mediante técnicas de alineamiento de secuencias [3] permite capturar de forma orgánica la variabilidad del lenguaje médico, sin la necesidad de intervención humana constante para la creación de diccionarios. Esta integración no solo tiene el potencial de mejorar el rendimiento del modelo, sino que también facilita la implementación de protocolos de explicabilidad mediante métodos como *SHapley Additive exPlanations* (SHAP) y *Contextual Importance and Utility* (CIU). Esto permite verificar si el sistema se basa en evidencia lingüística significativa o en artefactos estadísticos, transformando los clasificadores de alto rendimiento en herramientas técnicamente confiables y auditables para la revisión de expertos.

2. Estado del Arte

2.1. Evolución de la Clasificación de Textos Biomédicos

La clasificación automática de textos en el dominio biomédico es fundamental para la organización de información en registros de salud electrónicos y el apoyo a la toma de decisiones clínicas [2, 8].

2.1.1. De los Modelos Estadísticos a los Transformers

Históricamente, la disciplina dependía de métodos de aprendizaje automático tradicional, como *Support Vector Machines* (SVM) y *Naive Bayes* (NB), los cuales utilizan representaciones de "bolsa de palabras" (Bag of Words, BoW) para capturar frecuencias terminológicas. Sin embargo, la llegada de arquitecturas basadas en Transformers ha desplazado estos métodos debido a su capacidad para capturar dependencias contextuales de largo alcance. Modelos especializados como BioBERT [6, 12] han sido pre-entrenados en vastos corpus de literatura científica (PubMed), lo que les permite comprender la semántica técnica con una precisión superior a los modelos de propósito general.

2.1.2. El problema de la Transparencia y el Sesgo Estadístico

A pesar de su éxito predictivo, el despliegue de modelos de aprendizaje profundo en aplicaciones biomédicas enfrenta el desafío crítico de la transparencia [4]. Estos modelos operan a menudo como "cajas negras", dificultando la validación por parte de expertos clínicos y organismos regulatorios. Investigaciones recientes sugieren que confiar ciegamente en estas arquitecturas puede llevar a capturar sesgos o "ruido estadístico" en lugar de patrones lingüísticos con significado clínico real.

2.2. Marcos de Trabajo Híbrido y Reglas Lógicas

Para abordar la dicotomía entre rendimiento e interpretabilidad, ha emergido una tendencia hacia arquitecturas híbridas.

2.2.1. Automatización de Expresiones Regulares con CREGEX

Las expresiones regulares (RegExes) ofrecen transparencia intrínseca al representar patrones lingüísticos legibles por humanos. El algoritmo CREGEX permite automatizar la inducción de estas reglas directamente desde los datos de entrenamiento mediante técnicas de alineamiento de secuencias [3]. Este proceso elimina la necesidad de intervención manual intensiva y permite capturar de forma orgánica la variabilidad del lenguaje biomédico.

2.2.2. Integración por Fusión de Canales

La arquitectura híbrida moderna propone procesar el texto a través de dos canales paralelos: un canal de conocimiento explicito (basado en reglas lógicas) y un canal de conocimiento implícito (representaciones semánticas o embeddings). En modelos tradicionales,

esta integración se realiza mediante la concatenación horizontal de vectores de activación. En arquitecturas de aprendizaje profundo, la fusión ocurre en las capas finales, integrando *embeddings* contextuales con activaciones de reglas antes de la capa de decisión.

2.3. Inteligencia Artificial Explicable

La necesidad de modelos que sean no solo precisos sino también auditables ha impulsado el campo de la *eXplainable Artificial Intelligence* (XAI) [7, 11].

2.3.1. Métodos de Atribución Global: SHAP

El método SHAP (SHapley Additive exPlanations) se ha consolidado como una herramienta fundamental para cuantificar la contribución de cada característica (palabra o regla) a la predicción final [1, 9]. Mediante un enfoque aditivo basado en la teoría de juegos, SHAP permite identificar los términos con mayor peso en el modelo, facilitando una visión global de su comportamiento. No obstante, se ha observado que los métodos aditivos pueden tener limitaciones para capturar la especificidad de casos individuales.

2.3.2. Análisis Local y Utilidad Contextual: CIU

Como complemento a la visión global, el método CIU (Contextual Importance and Utility) permite evaluar la relevancia de una característica en una instancia particular. Ese enfoque distingue entre la Importancia Contextual (CI), que mide el peso de un término en la decisión, y la Utilidad Contextual (CU), que determina si dicho término apoya o contradice el diagnóstico predicho. El uso de CIU es especialmente efectivo en sistemas de soporte de decisión clínica, proporcionando una interpretabilidad mas intuitiva para el especialista [5].

3. Hipótesis y Objetivos

3.1. Hipótesis

Un framework híbrido, basado en la integración de algoritmos de *machine learning* con expresiones regulares generadas automáticamente, permitirá mejorar el desempeño y la interpretabilidad en tareas de clasificación de textos biomédicos, mediante la identificación de patrones lingüísticos relevantes a través de protocolos de explicabilidad y enmascaramiento. Bajo este escenario, se plantean las siguientes interrogantes:

- ¿Pueden las expresiones regulares generadas automáticamente mejorar el desempeño de los modelos de clasificación de texto biomédico?
- ¿Pueden las arquitecturas híbridas que combinan expresiones regulares y modelos de aprendizaje automático proporcionar predicciones interpretables mediante métodos de inteligencia artificial explicable?
- ¿Afecta significativamente el enmascaramiento de características lingüísticas relevantes para el dominio al rendimiento predictivo de los modelos propuestos?

3.2. Objetivo General

Desarrollar un framework híbrido y explicable para la clasificación de textos biomédicos, integrando algoritmos de machine learning con expresiones regulares generadas automáticamente, con el fin de obtener modelos técnicamente auditables e interpretables.

3.3. Objetivos Específicos

Para dar cumplimiento al objetivo general, se proponen los siguientes hitos de investigación:

1. Analizar el estado del arte en clasificación de textos biomédicos y técnicas de inteligencia artificial explicable, con énfasis en modelos híbridos basados en expresiones regulares y algoritmos de machine learning.
2. Crear conjuntos de datos biomédicos multilingües para tareas de clasificación de texto, considerando problemas binarios y multiclase.
3. Desarrollar un framework híbrido y explicable para la clasificación de textos biomédicos, integrando algoritmos de machine learning con expresiones regulares generadas automáticamente.
4. Evaluar el desempeño, interpretabilidad y robustez del framework propuesto mediante métricas de clasificación, técnicas de explicabilidad y experimentos de enmascaramiento.

4. Artículo

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

An Explainable Framework for Biomedical Text Classification Combining Regular Expressions and Machine Learning Models

MAURICIO A. FUENZALIDA[1], CHRISTOPHER A. FLORES[2] (Member, IEEE), and ROSA L. FIGUEROA [3]

[1]Department of Information Systems, Universidad del Bío-Bío, Concepción 4081112, Chile
[2]Department of Electrical and Electronic Engineering, Universidad del Bío-Bío, Concepción 4081112, Chile
[3]Department of Electrical Engineering, Universidad de Concepción, Concepción 4070409, Chile

Corresponding author: Christopher A. Flores (e-mail: cfloresj@ubiobio.cl).

This work was partly funded by UBB FAPEI FP2524107

ABSTRACT Text classification is essential for the automatic organization of biomedical information. However, its practical implementation remains challenging due to limited transparency and the need for expert validation. To address these issues, this work proposes an explainable hybrid framework for biomedical text classification that integrates machine learning models with interpretable logical rules based on Regular Expressions (RegExes). The main objective is to address the lack of transparency in classification algorithms, which often function as ''black boxes'' that are difficult to interpret. To this end, we design a framework that leverages two complementary sources of information: automatically generated patterns derived from RegExes with contextual and statistical representations derived from machine learning and Transformer-based models, such as Support Vector Machine (SVM), Naïve Bayes (NB), Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) and Sentence Transformer Fine-tuning (SetFit). To verify that the architecture learns meaningful biomedical patterns rather than merely statistical noise, an explainability protocol was applied using the SHapley Additive exPlanations (SHAP) and Contextual Importance and Utility (CIU) methods. In addition, masking tests were conducted in which key terms were hidden to measure how much the model's performance decreased without that essential information. Experiments conducted on three datasets, namely, the Chilean Waiting List Corpus (CWLC), the Medical Information Mart for Intensive Care III (MIMIC-III), and the Medical Abstracts Text Classification Dataset (MATC), demonstrate that the inclusion of logical rules improves the performance of traditional models by up to 4.5% in terms of F1-score. Furthermore, the explainability analysis revealed that masking domain-relevant terms produced performance drops of up to 50%, confirming that the proposed models rely on meaningful linguistic patterns rather than statistical noise. Thus, this work demonstrates that it is possible to develop high-performance classifiers that are both technically reliable and interpretable for expert review.

INDEX TERMS biomedical text classification, explainable artificial intelligence, regular expressions.

I. INTRODUCTION

NATURAL Language Processing (NLP) in the biomedical field has seen significant advances thanks to the availability of large volumes of data in electronic health records and scientific literature. The automatic classification of these texts is essential for organizing information and supporting clinical decision-making [1], [2]. In the Spanish-speaking world, the urgency of processing clinical narratives to manage the overload of healthcare systems has been highlighted, as in the case of waiting lists in Chile [3]. However, the deployment of these technologies in biomedical applications faces a critical challenge regarding the need for transparency and validation of the model's decisions by clinical experts and regulatory bodies [4].

Currently, the state-of-the-art is dominated by deep learning models, such as transformers, which achieve exceptional levels of accuracy. However, their ''black box'' nature makes it difficult to understand the internal mechanisms that lead to a specific classification, which may limit their adoption in scenarios where errors have serious clinical consequences

[5]. In contrast, Regular Expressions (RegExes) offer an inherently interpretable alternative, allowing linguistic patterns to be represented explicitly and in a way that is readable by humans. Despite their advantages in terms of transparency and computational efficiency, their manual design is a laborious task that requires in-depth knowledge of the domain and technical language [2].

To address this dichotomy between performance and interpretability, this work proposes an explainable hybrid framework for biomedical text classification. The proposed approach uses sequence alignment techniques to automatically extract representative patterns and generate RegExes from biomedical texts [2]. These patterns are integrated with classical machine learning models such as Naïve Bayes (NB) and Support Vector Machine (SVM), as well as Transformer-based models such as Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) and Sentence Transformer Fine-tuning (SetFit), combining explicit rule-based knowledge with statistical and contextual text representations [6], [7]. Recent advances in Transformer-based models, including architectures such as BioBERT and SetFit, have demonstrated strong performance in biomedical text classification tasks [6], [7]. However, despite their predictive capabilities, these approaches remain predominantly data-driven and often lack transparency in their decision-making processes. In this context, hybrid frameworks that integrate rule-based knowledge with machine learning models offer a promising approach to improve interpretability while preserving classification performance.

To evaluate the interpretability of the proposed framework, SHapley Additive exPlanations (SHAP) is used to identify the words and rules that most influence model predictions, while Contextual Importance and Utility (CIU) is applied to analyze the contextual importance (Contextual Importance (CI)) and utility (Contextual Utility (CU)) of each feature in the final decision-making process [8], [9]. A selective masking methodology is implemented to identify terms deemed critical by SHAP and to hide them from the original dataset, thereby quantitatively measuring the degradation in performance. This process allows to validate whether the models are basing their decisions on meaningful linguistic patterns or on artifacts in the dataset.

In this context, this work aims to address the following research questions:

- Can automatically generated RegExes improve the performance of biomedical text classification models?
- Can hybrid architectures combining RegExes and machine learning models provide interpretable predictions through Explainable Artificial Intelligence (XAI) methods?
- Does masking domain-relevant linguistic features significantly affect the predictive performance of the proposed models?

The rest of the article is organized as follows. Section II presents the state-of-the-art in biomedical Natural Language Processing (NLP) and XAI, focusing on the transition from traditional RegExes to transformer-based models and the emergence of hybrid interpretability frameworks such as SHAP and CIU. Section III details the methodology, including the processing of the three biomedical datasets and the architecture of the hybrid classifiers. Section IV presents the experimental results and the explainability analysis. Finally, Section V presents the conclusions and future work.

II. RELATED WORK

This section presents the state-of-the-art in biomedical text classification and XAI, focusing on the evolution from rule-based approaches to Transformer-based models, as well as on recent explainable hybrid frameworks that combine RegExes, machine learning models, and feature-attribution methods.

In recent years, transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) and its specialized variant BioBERT, have dominated clinical NLP tasks due to their ability to capture semantic relationships in large volumes of data. A recent systematic review shows that, between 2020 and 2024, the use of language models in healthcare has grown exponentially, with a clear shift from traditional methods toward the fine-tuning of specialized models [10]. In this regard, Hara *et al.* [11] have demonstrated that, for medical classification tasks, specific models from the BERT family typically outperform Large-Scale Generative Models (LLMs) due to their greater accuracy in technical contexts. Despite their high accuracy, these models operate as ''black boxes,'' which generates mistrust in medical settings where decision validation is critical for patient safety. Recent research suggests that blindly relying on these models may lead to the capture of statistical biases rather than actual linguistic patterns [5].

As an alternative to neural opacity, RegExes offer intrinsic transparency [12]. Various studies have sought to automate the creation of these patterns for text classification in the medical domain. For example, Cui *et al.* [13] propose a constructive heuristic to generate interpretable RegExes. In contrast, Tu *et al.* [14] extend this approach by using a simulated annealing algorithm to optimize rule quality without manual intervention. In this context, Sbei *et al.* propose a hybrid approach that uses RegExes for automatic sentence annotation, providing a cost-effective alternative to manual labeling; In their study, they demonstrated that distilled models such as DistilBERT can outperform much larger architectures and state-of-the-art generative models, such as GPT-4 and LLaMA-3, regarding accuracy on classification tasks, provided they have a well-defined rule base for initial labeling [15]. Following a hybrid approach, Li *et al.* [16] integrate RegExes with bidirectional LSTM networks equipped with attention mechanisms, aiming to balance accuracy and explainability. On the other hand, Puri and Patnaik [17] use sequence alignment techniques to derive patterns that serve as input for modified Support Vector Machine (mSVM)-based classifiers. This trend has recently been further explored through the use of Natural Language Inference models combined with RegExes for data extraction from oncology

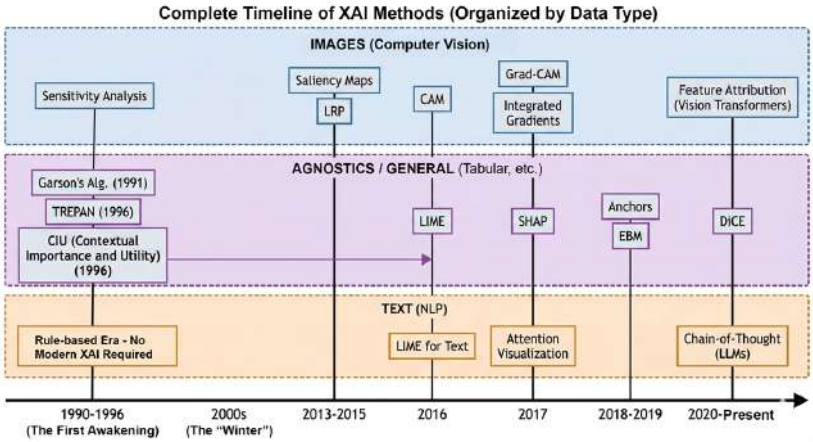

FIGURE 1. A comprehensive timeline of XAI methods. Categorized by data type and model architecture, highlighting the evolution from early methods such as CIU to today's LLMs.

reports, demonstrating that these hybrid architectures not only align closely with expert judgment but are also computationally lightweight and efficient for clinical research [18]. However, a limitation shared by these approaches is that the generated patterns tend to be rigid or, in the case of hybrid models, sacrifice some transparency by relying on complex neural architectures. To address these shortcomings, the Clinical Regular Expressions (CREGEX) algorithm [19] was developed, which enables the automatic generation of RegExes from training data. This approach not only preserves the transparency of the logical rules but also automates the capture of representative linguistic patterns, eliminating the need for intensive manual intervention. CREGEX [19] has automated the generation of RegExes using global and local alignment techniques such as Needleman-Wunsch and Smith-Waterman, enabling the induction of sequential patterns directly from text without constant human intervention.

Recent developments in clinical NLP have led to a trend toward hybrid models that seek to balance the predictive power of Transformers with the interpretability of logical rules [20]. In line with this trend, hybrid tools have been proposed that combine Transformer-based classifiers with classical rule-based systems such as NegEx to improve the detection of negations and temporal attributes in clinical records [21]. Likewise, the use of long-context encoders such as ModernBERT has begun to optimize the capture of dependencies in lengthy clinical notes [22]. At the same time, new auditing approaches suggest that explainability should be a cornerstone from the model's design stage [23]. In this context, auditing complex models has relied on various feature-attribution methods. As illustrated in Fig. 1, there is a wide range of XAI techniques that vary depending on the type of data and the level of access to the model; among these, techniques such as SHAP [8] have gained particular prominence. For example, Dolk *et al.* [24] apply SHAP to analyze prediction reliability in the automatic assignment of ICD-10 codes, while Dalhatu and Azmi Murad [25] use this technique to identify critical patterns in clinical narratives through feature selection. Likewise, its use has been explored in interpreting the severity of conditions in mental health consultations [26] and in developing explanations aimed at facilitating decision-making by healthcare professionals [27]. These applications facilitate the identification of terms or patterns that drive classification by using an additive approach to global importance.

However, after analyzing the options presented in the comparison in Fig. 1, it was found that additive methods have limitations in capturing the specificity of individual cases; therefore, the CIU method [9] was selected as a critical complement. For example, Främling [28] proposes this approach based on utility theory, which allows explanations to be generated without resorting to simplified or substitute interpretable models, thereby preserving the integrity of the original prediction from the black-box model. On the other hand, Knapič *et al.* [29] demonstrate that CIU is more effective in medical decision support systems, providing specialists with more intuitive interpretability than traditional additive attribution-based methods. Recently, Agrawal *et al.* proposed a benchmarking framework (XAI-Eval) to validate explainability methods in real-world clinical settings; their results confirm that, while methods such as Local Interpretable Model-Agnostic Explanation (LIME) are popular for local explanations, CIU offers superior performance in terms of user satisfaction and understanding of specific cases with high clinical sensitivity [30]. Sadeghi *et al.* also present a comprehensive review of XAI in healthcare, highlighting that the future of these technologies depends on the models' ability to handle the high dimensionality and noise inherent in clinical data. The authors emphasize the importance of moving toward explanations that are clinically meaningful and technically auditable, validating the trend toward architectures that prioritize interpretability without sacrificing predictive power [31]. Finally, Malhi and Främling [32] propose and validate this technique for the classification of complex clinical data, highlighting its ability to distinguish the importance of a term from its specific utility within a given diagnostic context.

A systematic review conducted by Noor *et al.* highlights that, despite the exponential growth of XAI in healthcare between 2000 and 2024, critical gaps remain, such as the lack of standardized evaluation metrics and the limited use of multiple combined explainability methods. This study identifies SHAP and LIME as the most prevalent techniques but recommends designing models focused on explainability and the complementary use of various XAI tools to overcome the trade-off between interpretability and fidelity in real-world clinical settings [33]. Unlike SHAP, CIU is based on Multi-Attribute Utility Theory and does not use a global approach; instead, it evaluates a feature's performance within a specific context. This distinction is critical in the clinical domain because while CI determines the weight of a clinical term in the decision, CU allows us to discern whether its presence supports or contradicts the predicted diagnosis.

Finally, it is important to indicate that our work differs from previous approaches, such as those by Flores *et al.* [19], [20], which primarily focus on classification performance based on RegExes, by proposing a hybrid two-channel architecture that combines RegExes and machine learning models. The proposed framework not only integrates explicit rule-based knowledge with statistical and contextual text representations but also incorporates an explainability-driven masking protocol to assess the models' reliance on meaningful linguistic patterns.

III. MATERIALS AND METHODS

This section details the proposed experimental design for the explainable classification of biomedical texts. The framework is divided into five main stages: (A) description and preprocessing of the datasets, (B) automatic generation of RegExes, (C) architecture of the hybrid classifiers, (D) explainability and contextual attribution, and (E) masking validation protocol.

A. DATASETS AND PRE-PROCESSING

Three biomedical text datasets were selected to evaluate the proposed framework across different linguistic and clinical contexts: the Chilean Waiting List Corpus (CWLC), the Medical Information Mart for Intensive Care III (MIMIC-III) Demo, and the Medical Abstracts Text Classification Dataset (MATC) [34]–[36]. These datasets include both clinical narratives and scientific biomedical abstracts in Spanish and English, allowing the evaluation of the proposed models under heterogeneous textual conditions.

We selected a subset of the CWLC, an open-access Spanish-language resource hosted on the Zenodo repository that specializes in Chilean clinical narratives [34]. Although the original corpus has a complex annotation structure with ten entity types, this study focused on a sample of approximately 1,000 records balanced across three specific classes (see Table 1).

On the other hand, the second corpus used was the MIMIC-III Demo [35], a representative English-language subset of the MIMIC-III, hosted on PhysioNet. This resource provides a controlled environment of de-identified real-world medical data. Preprocessing was aligned with the dataset's relational structure, focusing on information extraction while preserving the accuracy of the clinical notes.

Finally, the MATC [36] was incorporated, an English-language corpus of scientific literature abstracts designed to evaluate NLP models in the medical domain. This resource, originally hosted on GitHub and distributed via the Hugging Face platform, features a five-diagnostic-category structure. For this research, and in accordance with the methodology applied to the previous datasets, a subset of approximately 1,000 records distributed across three specific classes was selected (see Table 1). This selection allows the evaluation of the classifier's performance on highly formal, academic linguistic text, providing a necessary contrast to the clinical narrative of the other corpora used in this study.

TABLE 1. Description of the datasets used. This section details the sources and sizes of the subsets selected for the classification experiments, each limited to two or three specific classes.

Dataset	Number of classes	Number of samples
CWLC	3	1115
MIMIC-III Demo	2	2385
Medical Abstracts	3	1165

The texts underwent a preprocessing workflow designed to standardize lexical variants and reduce structural noise in the Spanish and English corpora. This process consisted, first, of character normalization through conversion to lowercase and Normalized Form Decomposition (NFD), which allowed for the systematic removal of diacritics and accents, ensuring consistent orthographic treatment regardless of language. Subsequently, tokenization based on RegExes was implemented, separating special characters and punctuation from word stems to facilitate independent analysis. To preserve the integrity of the quantitative data, a numerical reconstruction rule was applied, preventing the separation process from fragmenting decimal terms. Finally, class labels were normalized and encoded, ensuring a persistent data structure suitable for model training.

B. AUTOMATIC RULE GENERATION

Regular expression extraction is performed based on the CREGEX algorithm [19], which automates the induction of RegExes to transform unstructured clinical language into a formal feature space. Unlike traditional methods that rely on predefined dictionaries, this process autonomously identifies the most representative terms and patterns, structured into four technical phases:

1) Lexical analysis and automated vocabulary construction

The process begins with a text preparation stage in which the information is broken down into smaller units (tokenization), ensuring that punctuation marks do not interfere with the analysis and that important numerical data remains intact. The system automatically analyzes the entire document corpus to

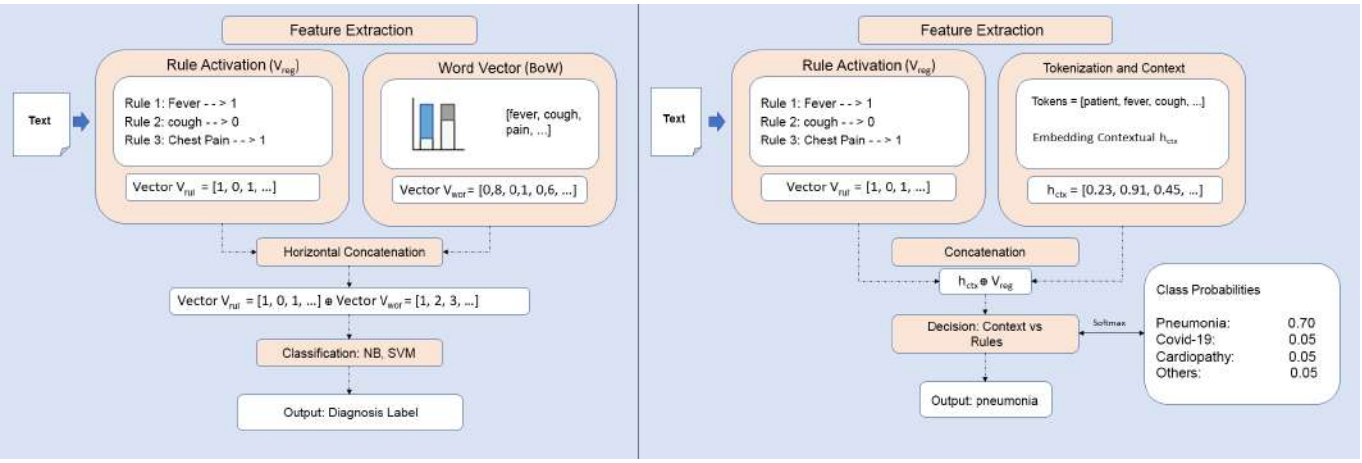

FIGURE 2. Hybrid integration architectures for biomedical text classification. Comparative representation of the proposed channel fusion mechanisms. On the left, the feature union approach for traditional models, where binary rule vectors (V_{reg}) and term frequencies (BoW) are horizontally concatenated. On the right, the layer fusion approach for transformer-based architectures, where the semantic summary or contextual embedding (h_{ctx}) is integrated with the rule vector via a concatenation operation ($\oplus$), prior to the final decision layer with Softmax activation.

identify the most common words and roots, without human intervention. During this analysis, terms that do not add medical value, such as conjugated verbs or administrative terms, are automatically removed, allowing the algorithm to focus solely on linguistically and semantically relevant concepts.

2) Adaptive sequence alignment

The core of the algorithm employs multiple sequence alignment techniques. The system automatically compares text fragments belonging to the same diagnostic category to detect linguistic constants and variables. This procedure allows for the organic capture of variability in medical language, identifying synonyms, verbal variations, and common spelling errors, and grouping terms with equivalent meanings under a single logical structure without the need for a user-provided list of keywords.

3) Pattern induction and generalization

Based on the alignments obtained, the system automatically generates RegExes. These patterns are not mere matches of static text; the algorithm introduces metacharacters and character classes to enable controlled generalization. This capacity for abstraction ensures that the rules identify recurring grammatical and clinical structures, allowing the system to inductively discover the most important descriptors for each class.

4) Feature selection and vector representation

Unlike the original CREGEX [19] approach, which focuses on the automatic generation of RegExes, the proposed framework incorporates a feature selection and vector representation stage designed to integrate rule-based patterns with machine learning models in an explainable hybrid architecture.

The feature selection process begins with filtering candidate rules to ensure the quality of the representation space. Each regular expression undergoes a syntactic validity check, discarding any pattern that triggers compilation errors. Subsequently, a minimum representativeness threshold of 2 characters based on the length of the token associated with the rule is applied: only those with an identifying label of two or more characters are integrated into the model, thereby eliminating structural noise and isolated symbols that do not provide relevant information.

Once validated, the rules are transformed into a representation vector called the ''Rule Channel''. This vector is constructed from a binary activation matrix, where each position indicates whether the pattern is present (1) or absent (0) in the processed document. This representation constitutes the core of the hybrid model, as it is concatenated with other feature vectors to expand the semantic context. In frequency-based models, this channel is integrated with a BoW matrix that captures the most frequent keywords extracted directly from the dataset after normalization and stopword removal. Meanwhile, in deep learning architectures, rule activations are fused with the language model's embedding vectors.

Finally, the evaluation of the discriminatory capacity of these features is not a manual preliminary step, but rather an analysis based on the relative importance that the model assigns to each vector component. The technical criterion used to audit which patterns were decisive in the prediction is the marginal impact calculated using SHAP values. This metric ranks both rules and keywords by their actual contribution to the classifier's output, providing an importance ranking based on dataset evidence rather than a static selection of terms.

C. HYBRID CLASSIFICATION ARCHITECTURE

The proposed architecture is based on a dual-representation approach that captures the synergy between explicit clinical knowledge and deep natural-language semantics. The system processes each text instance through two parallel information channels, the integration of which enables robust and auditable classification.

1) Explicit knowledge channel: rule-based

This workflow uses CREGEX [19] to generate a formal feature space. The text is evaluated against a dictionary of RegExes that extracts grammatical and lexical regularities specific to the biomedical domain. The result is a vector of binary activations $v_{rul} \in \{0,1\}^n$, where each component indicates the text's agreement with a specific logical rule.

2) Implicit knowledge channel: semantic representation

This channel extracts statistical and contextual information from biomedical texts using machine learning and Transformer-based representations. Depending on the experiment, either BoW representations are used to capture the most frequent terms in the corpus, or embedding-based models are employed to encode contextual semantic information. The fusion of both channels is performed using two distinct technical approaches, tailored to the nature of the final classifier:

- **Feature union approach:** For traditional classifiers, feature vectors are combined through horizontal concatenation. Rule activations are integrated with the BoW vector, allowing the model to evaluate explicit patterns and term-frequency information jointly.
- **Layer fusion approach:** In Transformer-based architectures, integration occurs through the concatenation of contextual embeddings and rule activations. The language model's contextual embedding is combined with the rule vector before the final decision layer, enabling the model to evaluate contextual semantics and explicit logical rules jointly.

As shown in Fig. 2, the proposed hybrid architecture processes each text instance through two complementary channels. The explicit knowledge channel (Channel 1) extracts the binary activation vector (V_{rul}) using logical rules. In contrast, the implicit knowledge channel (Channel 2) generates statistical or contextual representations depending on the classifier architecture. Both representations are subsequently concatenated, resulting in an enriched feature space for the final classification stage.

D. EXPLAINABILITY AND CONTEXTUAL ATTRIBUTION

To understand how the proposed hybrid models make decisions and verify that their predictions are supported by meaningful linguistic evidence, a dual explainability framework was implemented. This framework makes it possible to identify both the global importance of a feature and its contextual contribution within a specific prediction. SHAP values were used to quantify the contribution of each feature (word or rule) to the final prediction. In this study, SHAP is used not only to explain model behavior but also to identify the most influential features through masking experiments. These features are automatically selected according to their attribution scores, and their impact is evaluated by measuring the decrease in predictive performance after masking the corresponding linguistic information.

Mathematically, this method assumes that the final prediction is the sum of the individual attributions:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

According to [8], the term ϕ_0 represents the baseline value or the average prediction obtained from the entire training set. The specific impact of each feature i (whether a word or a CREGEX logical rule) is quantified using the SHAP values ϕ_i, which determine how much that element contributes to the deviation of the result from the average value. Finally, the binary variable $z'_i \in \{0,1\}^M$ allows us to model the presence or absence of each feature, facilitating the identification of the terms with the greatest weight to be selected in the masking experiment.

In addition, the CIU method is applied to assess the relevance of a feature in a particular case, avoiding the assumption that terms have a static weight across all scenarios. First, CI determines the impact a word or rule has on the current prediction, calculated by the difference in probability that the model achieves when varying that term between its best and worst theoretical possibilities:

$$CI = \frac{u_{max} - u_{min}}{U_{max} - U_{min}} \tag{2}$$

On the other hand, CU quantifies whether the presence of that characteristic actually supports the current prediction or generates uncertainty, placing the current prediction value $u(C)$ within the specific range of that variable:

$$CU = \frac{u(C) - u_{\min}}{u_{\max} - u_{\min}} \tag{3}$$

The incorporation of this analysis helps verify that the proposed models rely on meaningful linguistic patterns detected by CREGEX [19] and informative text representations rather than isolated statistical artifacts.

E. VALIDATION PROTOCOL BY MASKING

The validity of the generated explanations is evaluated through a masking experiment designed to determine whether the model relies on meaningful linguistic evidence or on statistical artifacts present in the dataset. This protocol is implemented differently depending on the classification architecture used, ensuring that the masking procedure remains technically consistent with each model's input representation and processing strategy.

For traditional models, the experiment focuses on the integrity of the input feature matrix. After obtaining the importance ranking via SHAP, the system selects up to 30 rules and 30 words with the highest cumulative impact for selective masking. The masking mechanism uses specific labels, employing `[MASK_W]` for lexical terms and `[MASK_R]` for matches detected by CREGEX patterns [19]. Once the original text has been modified, the script performs a complete re-vectorization using the `vectorizar_regex` and `CountVectorizer` functions, resulting in a feature matrix where rule activations and keyword frequencies are removed

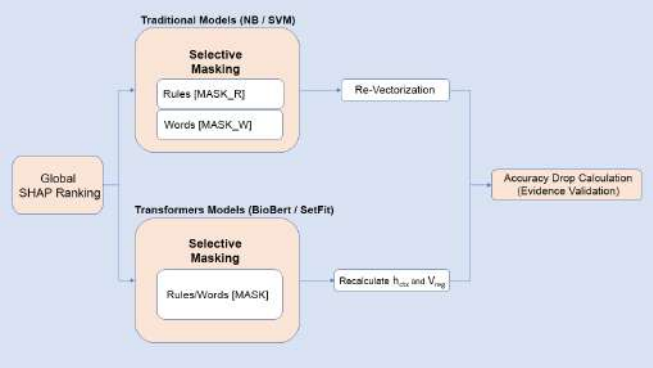

FIGURE 3. Flowchart of the masking validation protocol.

from the representation space. Sensitivity in this scenario is evaluated by measuring the decrease in predictive performance after masking these features.

On the other hand, in Transformer-based architectures, the protocol evaluates the robustness of the fusion mechanism and the contextual representation process. Here, masking is more extensive, selecting up to 30 rules and 30 high-impact words for neutralization. Unlike traditional feature-based approaches, the generic `[MASK]` token is used, which the transformer's tokenizer processes natively. When processing the masked text, the model generates a new contextual representation (h_{ctx}) that lacks information from the original linguistic patterns (e.g., words and rules). Simultaneously, the rule vector (V_{rul}) is dynamically recalculated by inserting zeros into the positions of the hidden rules. Validation in this model focuses on the degradation of predictive performance, allowing us to observe whether the model compensates for the loss of information using the remaining context or whether predictions strongly depend on the removed hybrid features.

Regardless of the model, the experiment's effectiveness is evaluated by the decrease in predictive performance observed after masking. A significant reduction in system performance indicates that the SHAP explanations accurately reflect the model's internal behavior and that the final classification relies on meaningful linguistic evidence. This protocol helps verify that the system depends on relevant information represented either explicitly through logical rules or implicitly through statistical and contextual text representations. While the masking protocol primarily validates the global feature attributions obtained with SHAP, CIU complements this analysis by providing contextual explanations at the individual prediction level. As shown in Fig. 3, this procedure ensures that the system remains auditable.

IV. RESULTS

This section presents the evaluation results of the proposed explainable hybrid framework, providing a comprehensive analysis of both predictive performance and interpretability. The analysis begins with an assessment of overall performance, comparing Accuracy (ACC) and F1-score (F1-score) metrics between the baseline models and their hybrid versions enriched with the automatic rule channel across the three evaluated datasets. Subsequently, feature attribution is analyzed using SHAP values to identify the linguistic features and logical patterns generated by CREGEX [19] that exert the greatest influence on the model's predictions. This analysis is complemented by the masking validation protocol, which quantifies the architectures' dependence on meaningful linguistic evidence and evaluates whether the models rely on relevant patterns or statistical artifacts. Finally, the section concludes with a local explainability analysis using the CIU method, which examines the contextual importance and utility of features in specific prediction cases, providing an interpretable view of the models' decision-making process within the proposed framework.

A. OVERALL PERFORMANCE EVALUATION

To validate the effectiveness of the proposed hybrid architecture and compare the performance of the baseline classifiers with their rule-channel-enhanced versions, standard classification evaluation metrics were used. The first of these is ACC, which measures the proportion of correct predictions relative to the total number of evaluated examples and is mathematically defined as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

In addition, the F1-score was used to obtain a metric that balances precision and recall, which is essential in the biomedical field to ensure the reliability of clinical decisions. This metric is calculated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{5}$$

In both equations, TP and TN correspond to correct classifications (True Positives and True Negatives), while FP and FN represent incorrect classifications (False Positives and False Negatives) [37].

The technical configuration and hyperparameters used to ensure the reproducibility of the results are detailed in Table 2. For the classical classifiers, a BoW vectorization was defined, limited to the 1000 most frequent features, and a minimum document frequency (min_df = 5) filter was applied to eliminate statistical noise and focus the analysis on terms with clinical relevance. Regarding Transformer-based models, such as BioBERT and SetFit, standardized configurations from the academic literature were employed, using a conservative learning rate of 2×10^{-5} and an AdamW optimizer to ensure stable convergence during training. The use of these hyperparameters, which are widely validated in NLP tasks [6], [38], ensures that the variations observed in feature importance through SHAP are not the result of arbitrary model tuning but instead faithfully reflect the architectural capabilities of each algorithm when facing the complexity of clinical language.

The results indicate that incorporating automatically generated logical rules consistently improves traditional models. On the CWLC dataset, NB saw a significant increase in ACC,

TABLE 2. Configuration and hyperparameters of the classification models.

Classifier	Architecture	Configuration and Hyperparameters
Naive Bayes	MultinomialNB	Vectorization: BoW; *max_features* = 1000; *min_df* = 5; Custom stopwords.
SVM	Linear SVC	Kernel: lineal; $C = 1.0$; BoW (*max_features* = 1000); Probability: enabled.
BioBERT	base-cased-v1.2	*Max_Len* = 128; *Batch* = 16; *Epochs* = 6; *LR* = 2×10^{-5}; Optimizer: AdamW; Dropout: 0.3.
SetFit	paraphrase-multil.	*Epochs* = 1; *LR* = 2×10^{-5}; *Batch* = 16; Loss: CosineSimilarityLoss; Base: MiniLM-L12-v2.

rising from 87.44% to 92.83% after integrating the explicit rule channel. A similar trend is observed in the MIMIC-III dataset, where all hybrid models achieved ACC levels above 98%, with BioBERT standing out with 99.16% accuracy after layer fusion. It is noteworthy that in the MATC dataset, the advantage of rules was most evident for NB (+4% improvement). To validate the consistency of these findings, the signed Wilcoxon rank-sum test for paired samples [37] was applied, confirming that the improvement in NB is systematic and not due to chance, yielding $p < 0.05$ across the three evaluated datasets. A similar trend was observed in the SVM model within the MATC dataset ($p = 0.033$), underscoring the effectiveness of hybridization in contexts with high terminological complexity.

B. SHAP VALUE ANALYSIS OF HYBRID MODELS

1) Analysis Using SHAP Values on the CWLC Dataset

The results presented below correspond to the attribution analysis of the CWLC dataset. This analysis reveals a distinction in the behavior of the hybrid models.

In the ''Enfermedad Cardiovascular'' (Cardiovascular Disease) class, the traditional regexNB and regexSVM models show a marked dependence on high-frequency terms; in the case of regexNB, the impact is concentrated on the word ''hipertension'' (hypertension) followed by the rule ?hiperten?[\s]** (hypertension-related pattern) (see Fig. 4). In contrast, the Transformer-based models assign higher attribution scores to more specific patterns such as (?:\w)? miocardiopatia(?:\w)? (cardiomyopathy-related pattern). Additionally, regexBioBERT highlights contextual terms such as ''cardiotoxicidad'' (cardiotoxicity), while regexSetFit assigns relevant importance to the comorbidity-related rule con[\s]*obesidad (with obesity), reflecting a more distributed attribution pattern across clinically related features.

For the ''Enfermedad Endocrina o Metabólica'' (Endocrine or Metabolic Disease) class, the divergence among architectures becomes more evident (see Fig. 5). The regexSVM model shows a strong attribution bias toward the word ''trastorno'' (disorder), followed by ''hipertension'' (hypertension), while regexNB maintains a similar distribution, prioritizing ''trastorno'' and ''hipotiroidismo'' (hypothyroidism). In contrast, regexBioBERT produces a more distributed attribution pattern by assigning high importance to both the ''trastorno'' rule (disorder-related pattern) and the word ''hipotiroidismo'' (see Fig. 6). Similarly, regexSetFit assigns relevant attribution scores to terms such as ''diabetes'' and the rule con[\s]*obesidad (with obesity), indicating a reduced dependence on generic high-frequency terms (see Fig. 7).

Finally, in the ''Trastorno Mental o del Comportamiento'' (Mental or Behavioral Disorder) class, the regexNB and regexSVM models rely predominantly on the word ''trastorno'' (disorder) and its regular expression variant *[\s]*trastorno (disorder-related pattern), particularly in regexSVM, where these features obtain substantially higher attribution scores than the remaining terms (see Fig. 5). In contrast, the Transformer-based hybrid models exhibit a more distributed attribution pattern. regexBioBERT identifies the ''trastorno'' rule as highly relevant but also assigns important attribution scores to contextual terms such as ''hipotiroidismo'' (hypothyroidism) and ''diabetes'', together with the symptom-related rule refiere[\s]*palpitaciones (reports palpitations) (see Fig. 6). Similarly, regexSetFit incorporates both the cardiovascular-related rule (?:\w)? miocardiopatia(?:\w)? (cardiomyopathy-related pattern) and mental health-related patterns into its importance ranking (see Fig. 7). Overall, these results suggest that Transformer-based hybrid models rely on a broader combination of contextual terms and automatically generated rules, whereas traditional models remain more dependent on isolated high-frequency keywords.

These differences in behavior suggest that while traditional models such as regexSVM exhibit a high sensitivity to the statistical repetition of terms, explaining why high-frequency words like ''trastorno'' (disorder) reach a cumulative impact near 55 points (see Fig. 5), Transformer-based models exhibit more distributed attribution patterns. By observing the axis scales, it becomes evident that in BioBERT and SetFit, importance is distributed across much lower ranges (between 8 and 13) (see Fig. 6, 7). This indicates that the attribution process is not influenced solely by how often a word appears, but also by its contribution to distinguishing between target classes. Thus, while regexNB classification is driven primarily by the frequency of isolated terms such as ''hipertension'' (hypertension), the Transformer-based models assign relevant importance to more specific patterns, such as (?:\w)? miocardiopatia(?:\w)? (cardiomyopathy-related pattern) (see Fig. 6, 7), despite not being the most frequent in the corpus. Ultimately, the SHAP analysis suggests that Transformer-based hybrid models reduce the influence of highly repetitive terms while assigning greater attribution to more contextually informative features.

2) Analysis Using SHAP Values in the MIMIC-III Dataset

The following section presents the analysis of the second dataset, known as MIMIC-III. Unlike the first dataset, this

TABLE 3. Classification results: Comparison between standalone and hybrid models and statistical significance analysis (*p*-value).

Classifier	CWLC			MIMIC-III			MATC		
	ACC	F1	*p*-val	ACC	F1	*p*-val	ACC	F1	*p*-val
NB	87.44%	87.42%	—	96.23%	96.23%	—	80.26%	80.34%	—
regexNB	**92.83%**	**92.85%**	0.011*	**98.32%**	**98.32%**	0.012*	**84.55%**	**84.53%**	0.012*
SVM	92.83%	92.83%	—	98.11%	98.11%	—	74.68%	74.67%	—
regexSVM	93.72%	93.69%	0.563	98.32%	98.32%	0.563	**71.67%**	**71.64%**	0.033*
BioBERT	96.41%	96.41%	—	98.32%	98.32%	—	88.41%	88.42%	—
regexBioBERT	95.96%	95.96%	0.317	99.16%	99.16%	0.157	86.27%	86.29%	1.000
SetFit	98.21%	98.20%	—	99.16%	99.16%	—	84.12%	84.06%	—
regexSetFit	98.21%	98.20%	—	99.16%	99.16%	—	83.69%	83.62%	0.563

Note: An asterisk (*) indicates statistical significance ($p < 0.05$). Significant differences were identified in the NB model (all datasets) and in the SVM (MATC).

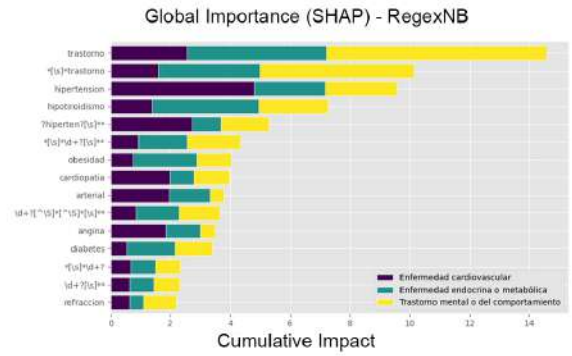

FIGURE 4. Feature Importance Distribution (SHAP) for the regexNB model.

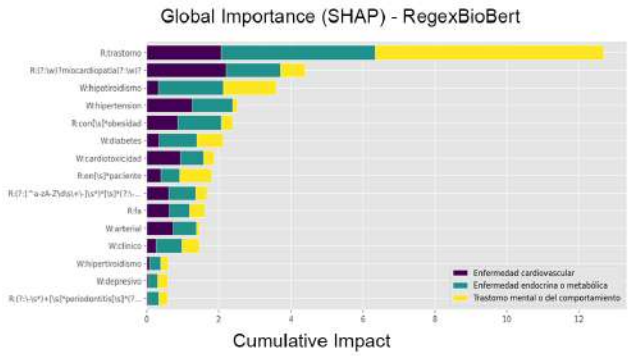

FIGURE 6. Feature Importance Distribution (SHAP) for the regexBioBERT model.

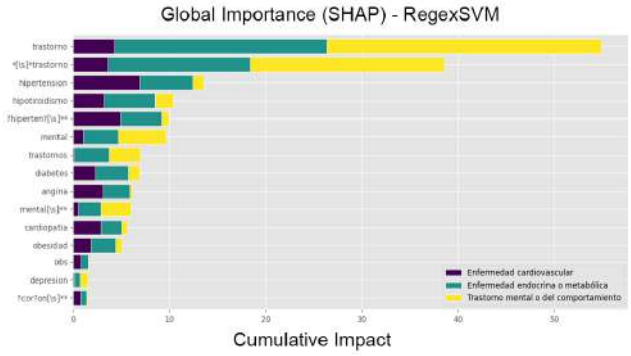

FIGURE 5. Feature Importance Distribution (SHAP) for the regexSVM model.

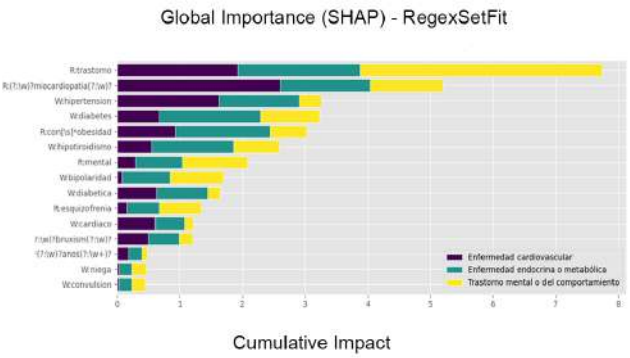

FIGURE 7. Feature Importance Distribution (SHAP) for the regexSetFit model.

dataset features a structure based on technical clinical language in English.

In the ''External Causes'' class, the regexNB and regexSVM models concentrate their highest attribution scores on high-frequency terms associated with incident descriptions. In regexNB, the strongest attribution is assigned to the word ''accident'', followed by ''history'' and ''status'' (see Fig. 8). Similarly, regexSVM assigns high importance to ''accident'', together with the rule *[\s]*by (causality-related pattern) and the term ''causing'', suggesting a stronger dependence on explicit causality-related keywords (see Fig. 9). In contrast, the Transformer-based models exhibit more distributed attribution patterns. regexBioBERT assigns high importance to the rules b and for[\s]*unspecified (unspecified-condition pattern), while also incorporating contextual terms such as ''screening'' (see Fig. 10). Likewise, regexSetFit highlights the ''effects'' rule together with the term ''injuring'', reflecting the integration of contextual terms and automatically generated rules within the attribution process (see Fig. 11).

Regarding the ''Supplementary Factors'' class, the results indicate that the models focus on identifying medical his-

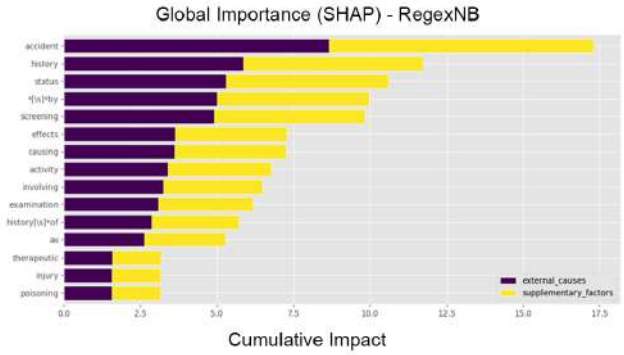

FIGURE 8. Feature Importance Distribution (SHAP) for the regexNB model.

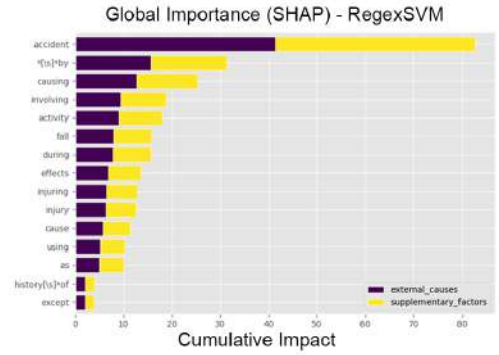

FIGURE 9. Feature Importance Distribution (SHAP) for the regexSVM model.

tory and current health status information. While traditional models such as regexNB and regexSVM rely primarily on frequent terms such as ''accident'', ''history'', and ''status'' for classification (see Figs. 8, 9), regexBioBERT exhibits a more detailed attribution pattern by emphasizing rules such as ''b'', for[\s]*unspecified (unspecified-condition pattern), and ''as''. Additionally, the following rule receives relevant attribution scores:

```
(?:\w)?l(?:\w)?ng[\s]*(?:\-\s*)+[\s]*(?:\w)?
    te(?:\w)?m[\s]*(?:[^a-zA-Z\d\s\+\-]\s*)
    *[\s]*current[\s]*(?:[^a-zA-Z\d\s\+\-]\s
    *)*[\s]*use[\s]*of
```

This pattern detects expressions associated with prolonged medication use (see Fig. 10). Similarly, regexSetFit assigns relevant importance to the ''effects'' rule, the word ''injuring'', and the rule (?:\w)?accident(?:\w+)? (accident-related pattern) (see Fig. 11). In both Transformer-based models, the words ''screening'' and ''history'' receive consistent attribution scores, suggesting that the hybrid framework captures contextual information associated with supplementary clinical factors in English-language texts.

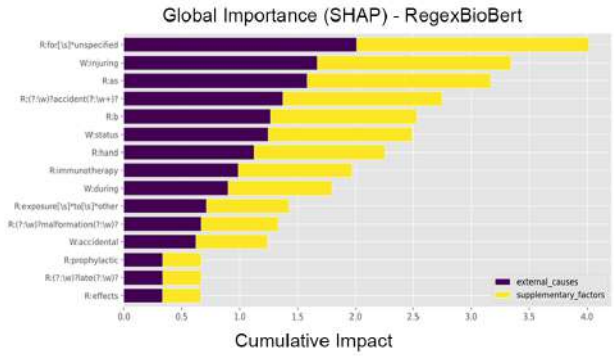

FIGURE 10. Feature Importance Distribution (SHAP) for the regexBioBERT model.

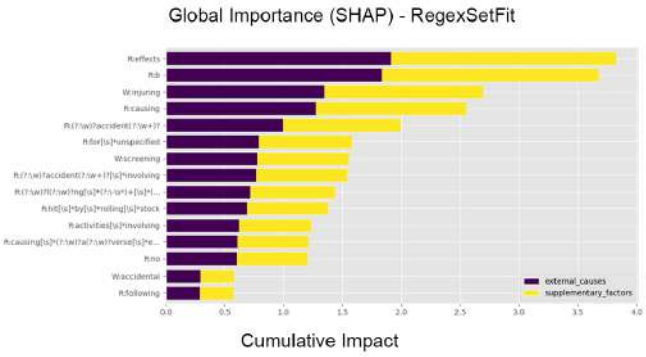

FIGURE 11. Feature Importance Distribution (SHAP) for the regexSetFit model.

3) Analysis Using SHAP Values on the MATC Dataset

The results analyzed in this section correspond to the third dataset, MATC. Unlike the previous datasets, this dataset consists of technical scientific literature, which requires the models to identify medical terminology with high precision.

In the ''Cardiovascular'' class, regexNB and regexSVM show a strong reliance on direct anatomical and pathological terms. In regexNB, the highest attribution scores are associated with words such as ''artery'', ''myocardial'', and ''hypertension'', whereas regexSVM emphasizes terms such as ''heart'' and ''percent'', reflecting a stronger dependence on cardiac function-related indicators (see Fig. 12). In contrast, the Transformer-based models assign relevant attribution scores to intervention- and molecular-related patterns. regexBioBERT highlights the rule ''radiotherapy'' together with the rule (?:\w)?ligand(?:\w)? (ligand-related pattern) and the keyword ''hypertension'' (see Fig. 14). Similarly, regexSetFit prioritizes the intervention-related rule (?:\w)?

revasculari(?:\w)?ation (revascularization-related pattern) together with the word ''hypertension'', reflecting the integration of contextual terms and automatically generated rules within the attribution process (see Fig. 15).

Regarding the ''Digestive System'' class, the attribution analysis reveals a strong focus on organ-specific and inflammation-related terms (see Fig. 13). The regexNB and regexSVM models concentrate their highest attribution scores on terms such as ''liver'', ''hepatitis'', ''abdominal'', and ''carcinoma''. In contrast, regexBioBERT assigns high importance to the rules (?:\w)?lower(?:\w)?[\s]*gastrointestinal (lower gastrointestinal-related pattern) and (?:\w)?frequen(?:\w+)?[\s]*hepatic (hepatic-related pattern), while also incorporating the term ''abdominal'' as a relevant contextual feature. Similarly, regexSetFit prioritizes the rule (?:\w)?frequen(?:\w+)?[\s]*hepatic together with the word ''liver'', and additionally identifies procedure-related terms such as ''laparoscopic'' (see Fig. 14). Overall, these attribution patterns suggest that the hybrid models integrate contextual medical terminology and automatically generated rules to represent digestive-system-related conditions.

Finally, in the ''Neoplasms'' classification, the attribution patterns focus primarily on cellular biology and cancer-treatment-related terminology. regexNB and regexSVM rely predominantly on frequent terms such as ''cell'', ''carcinoma'', ''tumor'', and ''cancer''. In contrast, the Transformer-based models exhibit more distributed attribution patterns. regexBioBERT assigns the highest attribution scores to the ''radiotherapy'' rule, followed by more specific patterns such as (?:\w)?cholangiocarcinoma(?:\w)?(cholangiocarcinoma-related pattern) and (?:\w)?lower(?:\w)?[\s]*gastrointestinal (lower gastrointestinal-related pattern) (see Fig. 14). Similarly, regexSetFit assigns greater importance to terms such as ''liver'' and ''bowel'' in combination with the ''radiation'' rule. Overall, these results suggest that Transformer-based hybrid models incorporate both contextual oncological terminology and automatically generated rules into the attribution process, thereby reducing dependence on isolated high-frequency words.

These attribution patterns reveal substantial differences in how importance is distributed across architectures in the MATC dataset. In the regexSVM model (see Fig. 13), the variable ''tumor'' reaches an attribution value close to 9.7 points on a scale extending to 10, suggesting a strong dependence on highly frequent keywords within the ''Neoplasms'' class. A similar trend is observed in regexNB (see Fig. 12), where the term ''liver'' reaches an attribution value near 6.8 points on a scale of 7, concentrating a substantial portion of the attribution on a single high-frequency term.

In contrast, the Transformer-based models exhibit more distributed attribution patterns. In regexBioBERT (see Fig. 14), the maximum attribution values remain within a substantially lower range, reaching approximately 2.6 points and led by the contextual rule ''R:radiotherapy''. Similarly, in regexSetFit (see Fig. 15), the highest attribution score corresponds to a regex pattern associated with hepatic frequency,

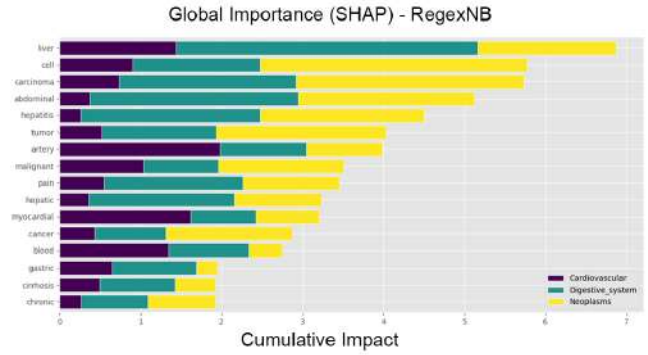

FIGURE 12. Feature importance distribution (SHAP) for the regexNB model.

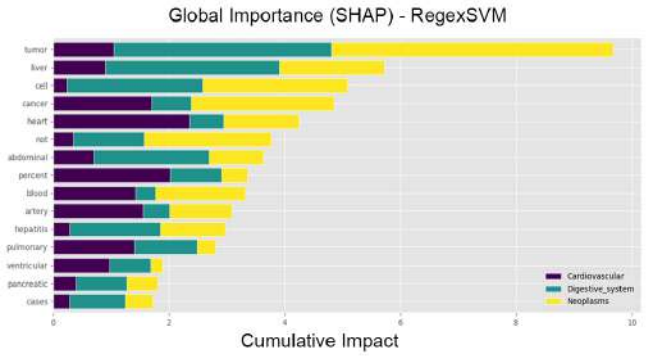

FIGURE 13. Feature importance distribution (SHAP) for the regexSVM model.

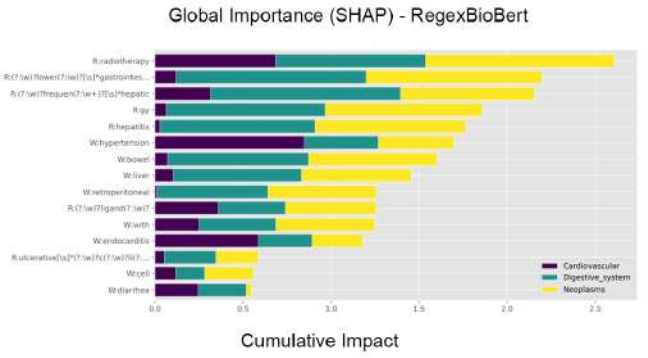

FIGURE 14. Feature importance distribution (SHAP) for the regexBioBERT model.

with values close to 2.2 points. These observations suggest that while traditional models rely more strongly on highly repetitive keywords, Transformer-based models distribute attribution across multiple contextual terms and automatically generated rules, reducing the dependence on isolated high-frequency features.

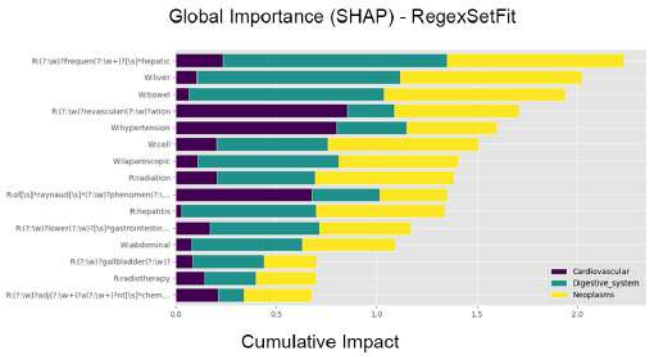

FIGURE 15. **Feature importance distribution (SHAP) for the regexSetFit model.**

C. MASKING-BASED FEATURE IMPORTANCE ANALYSIS

To evaluate the reliability of the attribution process, masking experiments were performed by selectively removing the most influential features identified through SHAP. This analysis makes it possible to quantify the dependence of each architecture on explicit rules, contextual terms, and highly attributed linguistic patterns across the evaluated datasets.

1) Analysis on the CWLC Dataset

In the CWLC dataset, the regexNB model drops from 92.82% to 78.92% ACC after masking (see Fig. 16), representing a reduction close to 14 percentage points. This decline suggests that regexNB relies substantially on the masked features, although it preserves part of its predictive capability through the remaining lexical information.

The regexSVM model exhibits the highest sensitivity to masking, decreasing from 93.72% to 35.87% ACC, corresponding to a 61.72% reduction (see Fig. 16). This substantial decline indicates a strong dependence on the explicit presence of highly attributed words and rules identified through the SHAP analysis.

In contrast, the Transformer-based architectures preserve a higher proportion of their predictive performance after masking. regexBioBERT decreases from 95.51% to 73.54% ACC, while regexSetFit drops from 98.21% to 62.78% (see Fig. 16). Although both architectures remain sensitive to the removal of highly attributed features, their residual performance suggests a greater capacity to preserve contextual information after masking.

2) Analysis on the MIMIC-III Dataset

In the MIMIC-III dataset, the masking experiment reveals substantially greater robustness across all architectures. regexNB experiences a reduction from 98.32% to 88.05% ACC, corresponding to a 10.45% decline (see Fig. 18). This behavior suggests that the model continues to benefit from the remaining contextual and lexical information despite the removal of highly attributed features.

regexSVM again exhibits the strongest sensitivity to masking. After removing the top attributed features, performance

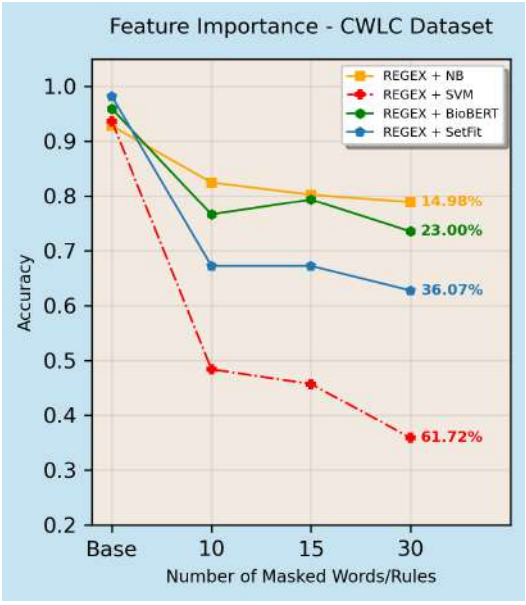

FIGURE 16. **Impact of feature masking on model accuracy in the CWLC dataset.**

decreases from 98.32% to 52.62% ACC, representing a 46.48% decline (see Fig. 18). This substantial reduction indicates a strong dependence on explicitly identifiable words and rule-based patterns.

In contrast, regexBioBERT demonstrates the highest robustness to feature masking, recording only a 7.02% reduction after masking. Similarly, regexSetFit preserves relatively stable performance despite the removal of highly attributed features. These observations suggest that Transformer-based architectures preserve a greater amount of contextual information after masking compared with traditional classifiers.

3) Analysis on the MATC Dataset

The masking analysis on the MATC dataset reveals different levels of dependence on highly attributed terminology across architectures. regexNB decreases from 84.55% to 78.54% ACC, corresponding to a 7.11% reduction after masking (see Fig. 20). This relatively moderate decline suggests that the model retains part of its predictive capability through the remaining specialized vocabulary present in scientific abstracts.

As observed in the previous datasets, regexSVM exhibits the highest sensitivity to masking. Performance decreases from 71.67% to approximately 52% ACC, representing a 27.54% reduction (see Fig. 20). This behavior indicates that the classifier relies strongly on highly attributed terms and automatically generated rules for distinguishing between the evaluated categories.

In contrast, regexBioBERT and regexSetFit demonstrate the greatest robustness to masking, recording reductions of 6.15% and 4.62%, respectively. These results suggest that Transformer-based architectures better preserve contextual semantic information after the removal of highly attributed features, reducing the dependence on isolated high-frequency

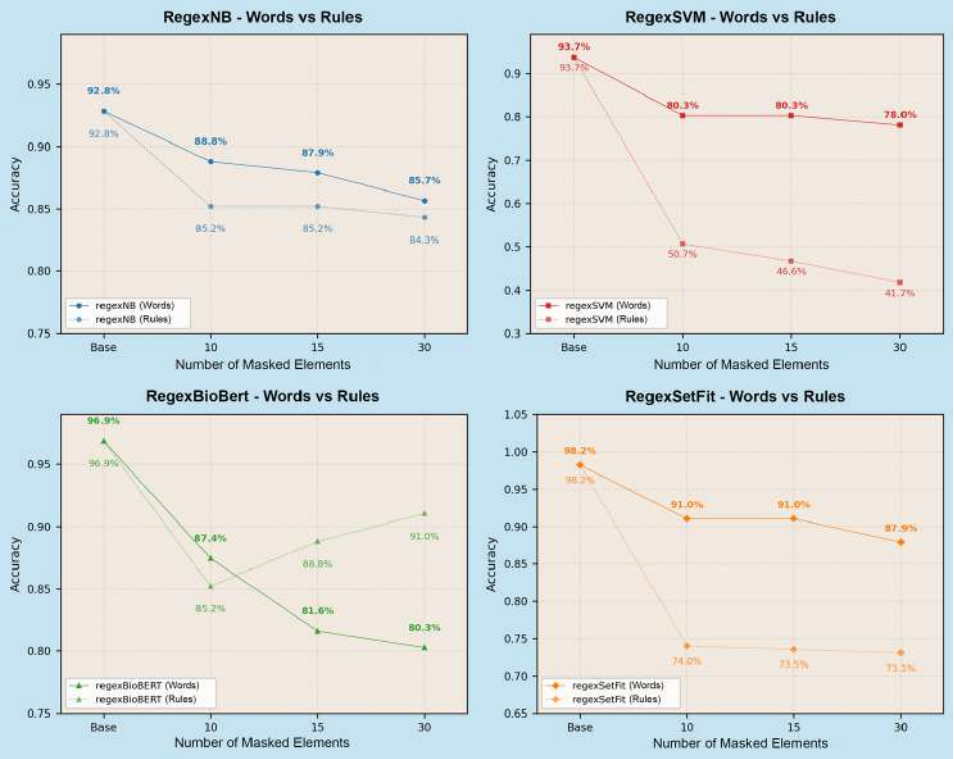

FIGURE 17. **Impact of word (W) and rule (R) masking on model accuracy in the CWLC dataset.**

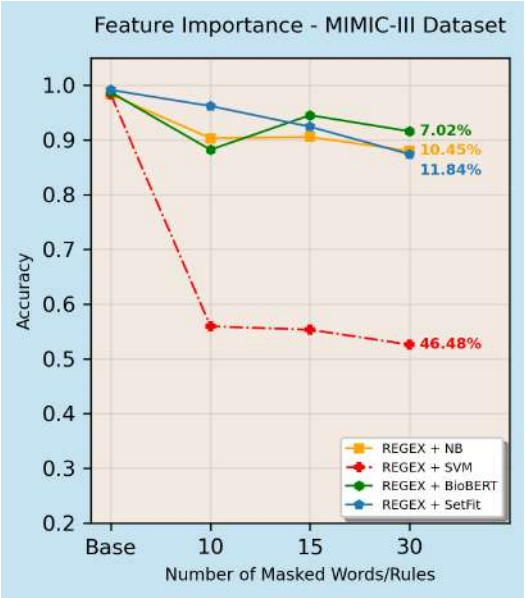

FIGURE 18. **Impact of feature masking on model accuracy in the MIMIC-III dataset.**

terms.

D. COMPARATIVE ANALYSIS: RULE-BASED VS. WORD-BASED MASKING

This analysis evaluates the relative contribution of words and automatically generated rules by progressively masking each feature type independently. The objective is to determine how strongly each architecture depends on lexical information and explicit rule-based representations across the evaluated datasets.

1) Analysis on the CWLC Dataset

When analyzing sensitivity to word masking, all models exhibit progressive performance degradation as more highly attributed terms are removed. regexBioBERT shows the largest reduction after masking 30 words, indicating a strong dependence on contextual lexical information. In contrast, regexNB preserves comparatively stable performance, suggesting that its probabilistic structure distributes attribution across a broader set of lexical features. regexSVM and regexSetFit exhibit intermediate behavior, maintaining partial robustness despite the progressive removal of highly attributed words.

The analysis of rule masking reveals substantially different behavior across architectures. regexSVM exhibits the strongest dependence on automatically generated rules, with performance decreasing from 93.72% to 41.70% after masking 30 rules. Similarly, regexSetFit demonstrates notable sensitivity to rule masking, indicating that explicit rule-based representations contribute substantially to its predictive performance. In contrast, regexBioBERT preserves comparatively stable performance after the initial degradation, suggesting that contextual semantic representations partially compensate for the removal of explicit rules.

Overall, these observations indicate that hybridization contributes differently across architectures. While regexSVM

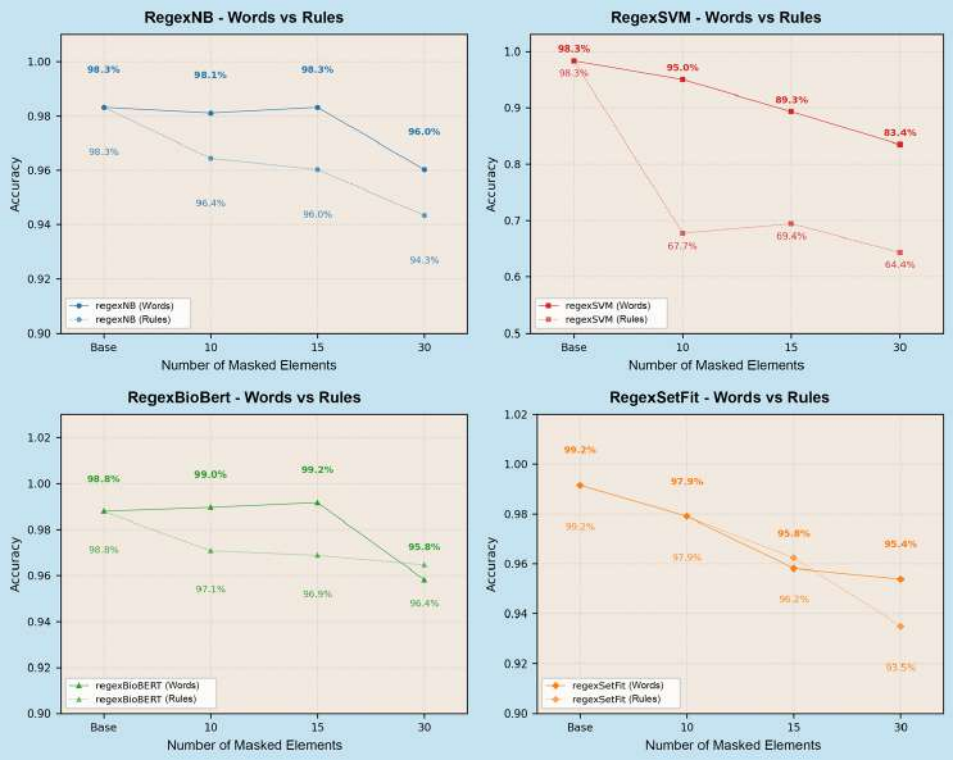

FIGURE 19. **Impact of word (W) and rule (R) masking on model accuracy in the MIMIC-III dataset.**

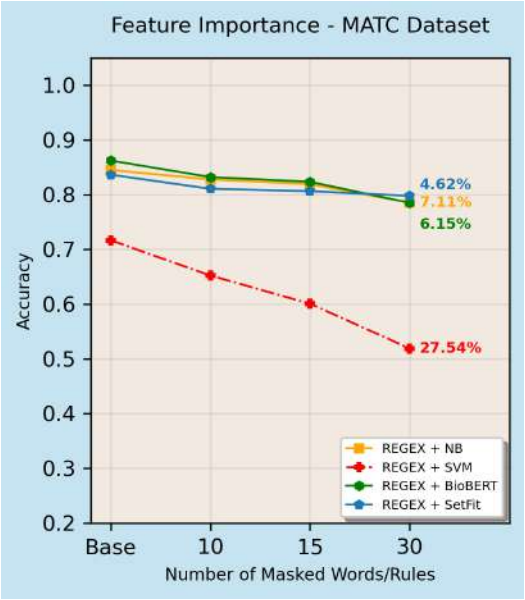

FIGURE 20. **Impact of feature masking on model accuracy in the MATC dataset.**

and regexSetFit rely more strongly on explicit rule-based evidence, regexBioBERT preserves a greater proportion of its contextual information after masking. regexNB exhibits the most balanced degradation between lexical and rule-based masking conditions.

2) Analysis on the MIMIC-III Dataset

When evaluating sensitivity to word masking, the Transformer-based architectures demonstrate substantially greater robustness. regexBioBERT maintains nearly stable performance during the initial masking stages and exhibits only moderate degradation after masking 30 words. Similarly, regexSetFit preserves high accuracy throughout the masking process, suggesting that contextual semantic representations mitigate the loss of isolated lexical features. In contrast, regexSVM shows the greatest degradation after word masking, suggesting a greater reliance on explicit lexical information.

The analysis of rule masking again reveals that regexSVM exhibits the strongest structural dependence on automatically generated rules. After masking 30 rules, the model shows a substantial performance drop, confirming the importance of explicit rule-based features for classification. In contrast, regexBioBERT and regexSetFit remain comparatively stable under rule masking conditions, preserving performance above 90% throughout most masking levels.

Overall, the results obtained on the MIMIC-III dataset suggest that Transformer-based architectures rely less on isolated lexical or rule-based features and instead preserve contextual semantic information more effectively after masking. In contrast, regexSVM remains strongly dependent on explicit linguistic evidence represented through both words and rules.

3) Analysis on the MATC Dataset

In the scientific literature, as represented by MATC, Transformer-based architectures again demonstrate the high-

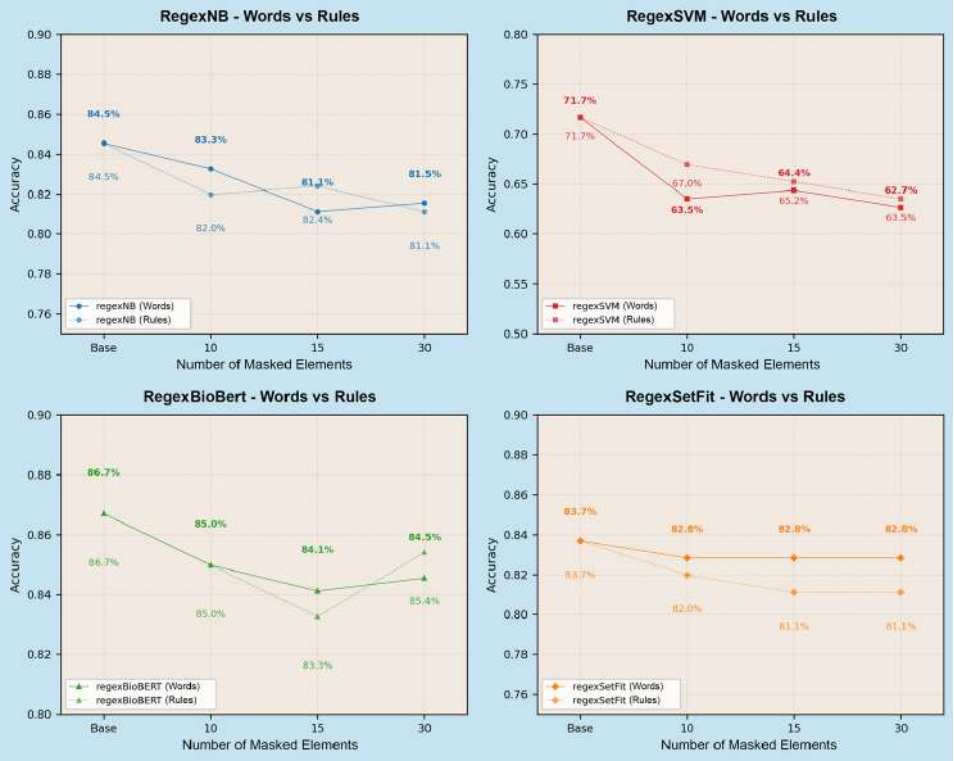

FIGURE 21. Impact of word (W) and rule (R) masking on model accuracy in the MATC dataset.

est robustness to word masking. regexBioBERT experiences only moderate reductions after masking highly attributed terms, while regexSetFit preserves stable performance across all masking levels. These observations suggest that contextual semantic representations remain effective even after removing important technical terminology.

In contrast, regexSVM exhibits the strongest lexical degradation, indicating a greater reliance on highly attributed technical terms for classification in scientific abstracts. regexNB demonstrates intermediate behavior, preserving part of its predictive capability despite the progressive removal of lexical features.

The rule-masking analysis reveals a more distributed dependence on automatically generated rules than in previous datasets. Although regexSVM remains the architecture most affected by rule masking, the performance reduction is smaller than that observed in CWLC and MIMIC-III. This behavior suggests that, in the scientific literature, contextual terminology itself provides substantial discriminative information, reducing the relative contribution of explicit rule-based features.

Overall, the MATC results indicate that as the textual domain becomes more technical and specialized, Transformer-based architectures preserve a clearer advantage in robustness to masking. These models maintain comparatively stable performance through the integration of contextual semantic information and automatically generated rules, whereas traditional architectures remain more dependent on isolated lexical and rule-based features.

E. LOCAL EXPLANATORY ANALYSIS USING CIU

As a final interpretability stage, the CIU method was implemented to evaluate the contextual importance and utility of linguistic features within individual predictions. Unlike the global attribution analysis provided by SHAP, CIU allows the contribution of words and automatically generated rules to be analyzed at the instance level.

- **Contextual Importance (CI):** Measures the influence of a feature on the model prediction within the analyzed instance.
- **Contextual Utility (CU):** Measures the extent to which a feature supports or contradicts the predicted class in the analyzed context.

The analyzed cases were selected as representative examples of challenging prediction scenarios across the evaluated datasets. Specifically, the selected instances include systematic misclassifications, ambiguous contextual patterns, and cases where the models exhibit conflicting contextual importance and utility distributions. This selection allows the CIU analysis to illustrate how different architectures prioritize contextual terms and automatically generated rules at the local prediction level.

1) Analysis Using CIU Values on the CWLC Dataset (Case 138)

The local explainability analysis for Case 138 reveals a shared misclassification pattern across the evaluated architectures. Although the correct category corresponds to ‘‘Enfermedad endocrina o metabólica’’ (Endocrine or Metabolic Disease),

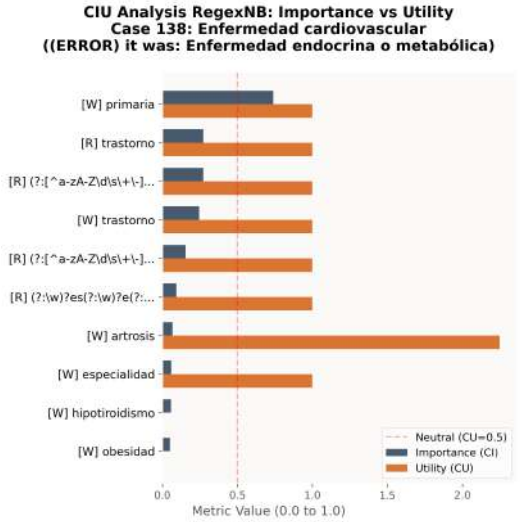

FIGURE 22. Local explainability analysis using the CIU method (CWLC dataset - regexNB).

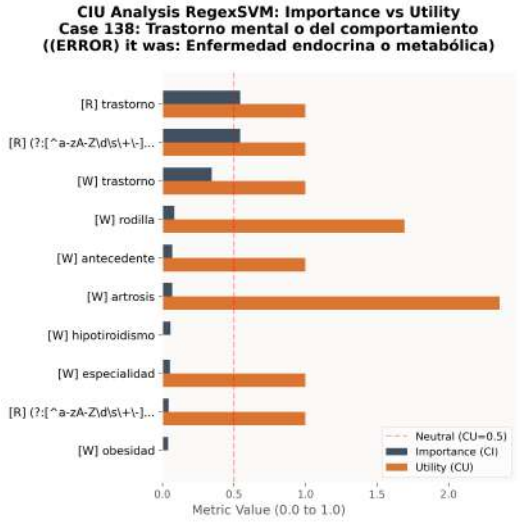

FIGURE 23. Local explainability analysis using the CIU method (CWLC dataset - regexSVM).

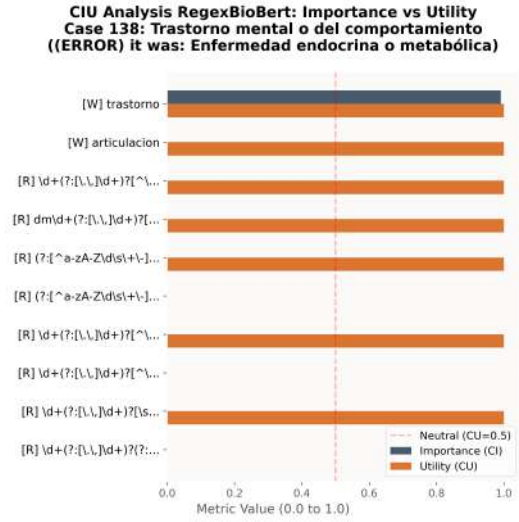

FIGURE 24. Local explainability analysis using the CIU method (CWLC dataset - regexBioBERT).

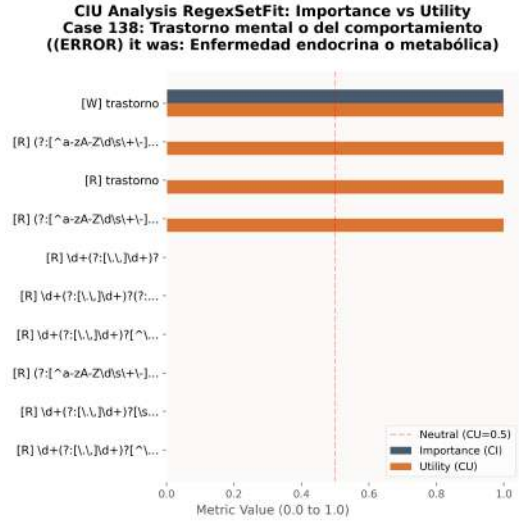

FIGURE 25. Local explainability analysis using the CIU method (CWLC dataset - regexSetFit).

the models assign high contextual importance to generic clinical terms associated with alternative categories.

The regexNB classifier exhibits a relatively dispersed attribution distribution but incorrectly predicts ''Enfermedad cardiovascular'' (Cardiovascular Disease). The term ''primaria'' (primary) receives the highest contextual importance ($CI \approx 0.75$), becoming the dominant contributor to the prediction. In contrast, more specific metabolic terms such as ''hipotiroidismo'' (hypothyroidism) and ''obesidad'' (obesity) exhibit comparatively low importance values despite their contextual relevance.

The regexSVM model predicts ''Trastorno mental o del comportamiento'' (Mental or Behavioral Disorder), exhibiting a strong lexical dependence on the word ''trastorno'' (disorder) and its associated regular expression variants. These features dominate the contextual importance metric ($CI \approx 0.55$), while terms such as ''artrosis'' (arthrosis) and ''rodilla'' (knee) present high contextual utility but comparatively low importance values. This behavior suggests that the classifier prioritizes highly frequent lexical patterns over more specific metabolic terminology.

Similarly, regexBioBERT and regexSetFit exhibit highly concentrated attribution patterns centered on the term ''trastorno''. In both architectures, this term reaches contextual importance and utility values close to 1.0, indicating a strong association between the feature and the predicted category. Although Transformer-based architectures generally demonstrate greater contextual robustness in the global analyses, this case illustrates that highly dominant lexical patterns may still influence local predictions disproportionately.

Overall, the CIU analysis for Case 138 reveals that all architectures assign greater importance to generic

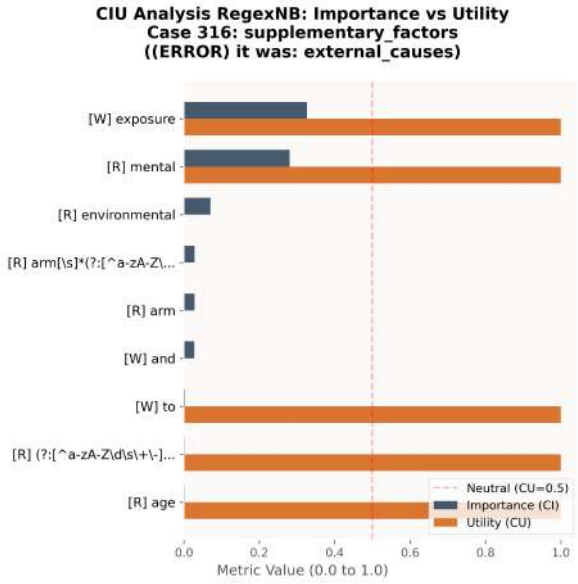

FIGURE 26. Local explainability analysis using the CIU method (MIMIC-III dataset - regexNB).

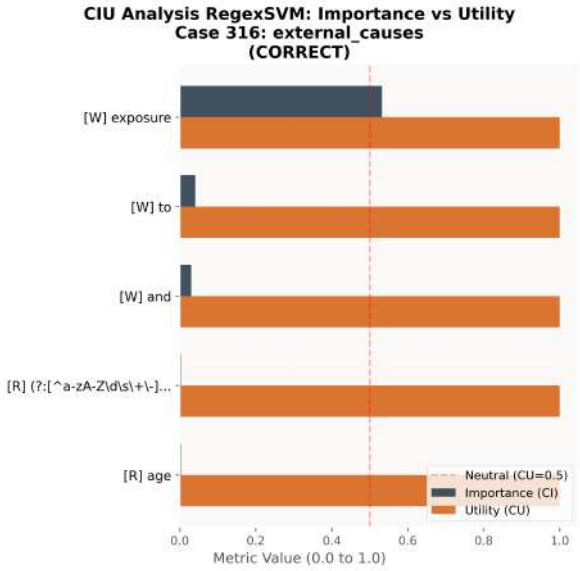

FIGURE 27. Local explainability analysis using the CIU method (MIMIC-III dataset - regexSVM).

high-frequency clinical descriptors than to more specific endocrine-related terminology. This behavior helps explain the shared misclassification observed across the evaluated models.

2) Analysis Using CIU Values on the MIMIC-III Dataset (Case 316)

The analysis of Case 316 reveals a substantially more stable classification scenario across the evaluated architectures. In this case, the correct category corresponds to ''external_causes'', which is associated with environmental factors and accident-related descriptions.

The regexNB classifier is the only architecture that incorrectly predicts ''supplementary_factors''. The model assigns moderate contextual importance to terms such as ''exposure'' and ''mental'', while the latter acts as a distractor feature with $CI \approx 0.3$. This attribution pattern suggests that the probabilistic structure of regexNB partially associates environmental exposure terms with supplementary clinical information.

In contrast, regexSVM correctly identifies the ''external_causes'' category by assigning dominant contextual importance to the term ''exposure'', which reaches high *CI* and *CU* values. Less informative terms such as ''to'' and ''and'' retain low contextual importance despite their presence in the text, indicating that the model filters part of the lexical noise during prediction.

The Transformer-based architectures exhibit comparatively more distributed contextual attribution patterns. In regexBioBERT and regexSetFit, the importance of individual features remains comparatively low, while the utility values associated with contextual patterns such as ''arm'' and environmental-related regular expressions remain high. These observations suggest that the Transformer-based architectures rely more strongly on the combined contextual

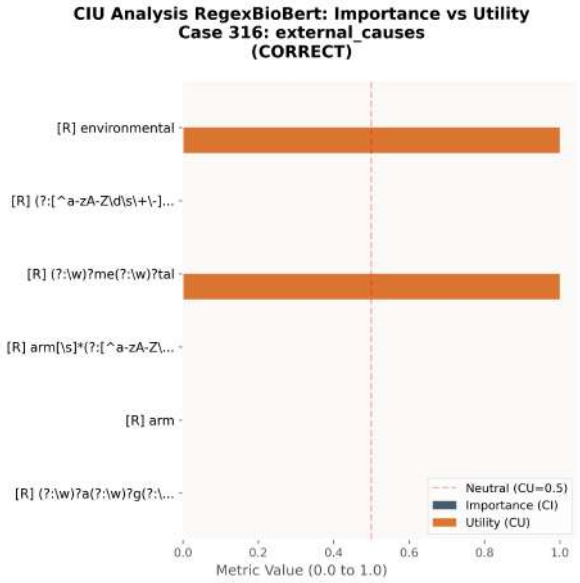

FIGURE 28. Local explainability analysis using the CIU method (MIMIC-III dataset - regexBioBERT).

representation of multiple features rather than on isolated keywords.

Overall, the CIU analysis on the MIMIC-III dataset demonstrates that Transformer-based architectures preserve more stable contextual attribution patterns under ambiguous lexical conditions, whereas regexNB remains more sensitive to distractor terms.

3) Analysis Using CIU Values on the MATC Dataset (Case 188)

The analysis of Case 188 reveals a shared semantic conflict across all evaluated architectures. Although the correct

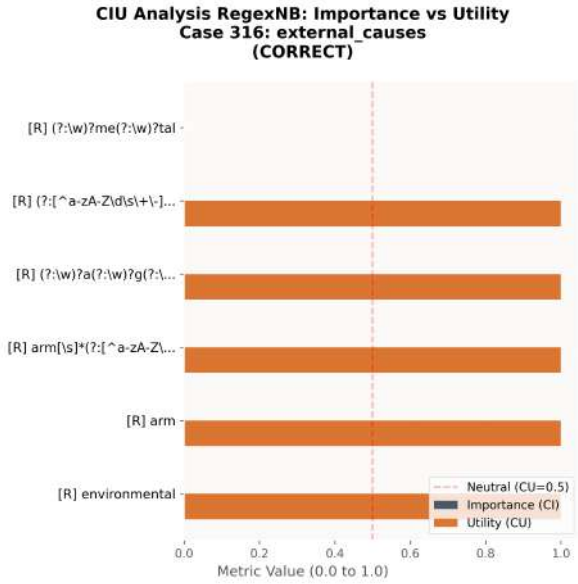

FIGURE 29. Local explainability analysis using the CIU method (MIMIC-III dataset - regexSetFit).

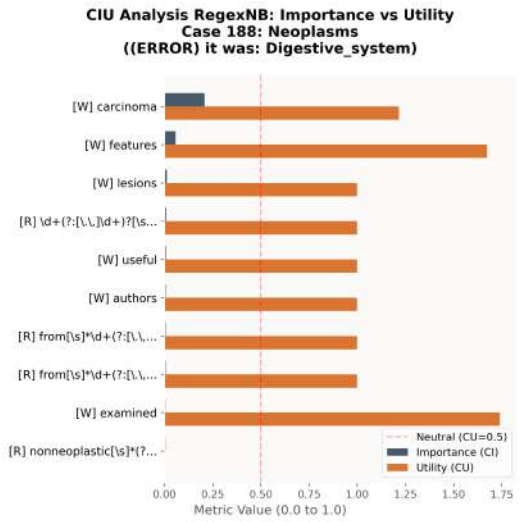

FIGURE 30. Local explainability analysis using the CIU method (MATC dataset - regexNB).

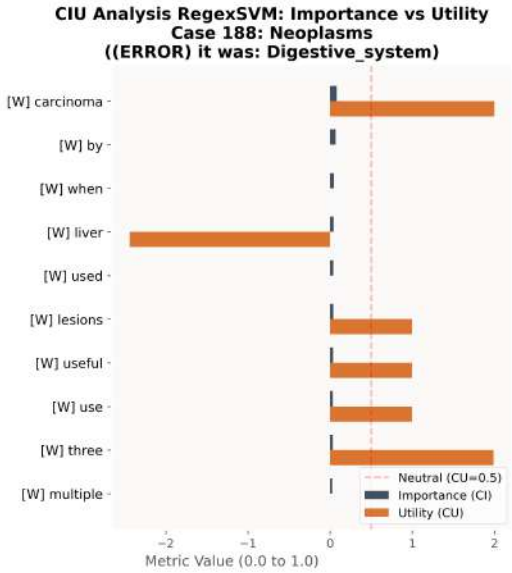

FIGURE 31. Local explainability analysis using the CIU method (MATC dataset - regexSVM).

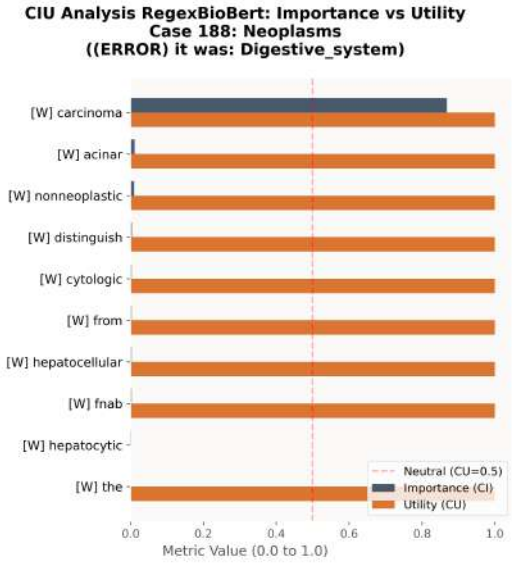

FIGURE 32. Local explainability analysis using the CIU method (MATC dataset - regexBioBERT).

category corresponds to ''Digestive_system'', all classifiers incorrectly predict ''Neoplasms''. This behavior is strongly associated with the contextual importance assigned to the term ''carcinoma'', which dominates the attribution process across models.

In regexBioBERT and regexSetFit, the contextual importance assigned to ''carcinoma'' reaches values close to 0.9 and 1.0, respectively, indicating a dominant association between oncological terminology and the predicted class. This behavior reduces the relative contribution of anatomical terms associated with the digestive system.

A particularly relevant behavior is observed in regexSVM, where the term ''liver'' exhibits strongly negative contextual utility values ($CU \approx -2.5$). Although this feature contradicts the predicted ''Neoplasms'' category, its contextual importance remains comparatively low and therefore does not reverse the final prediction. Similarly, regexNB distributes contextual utility across generic scientific terms such as ''features'' and ''examined'', while still assigning dominant importance to oncological terminology.

Overall, the CIU analysis on the MATC dataset suggests that the evaluated architectures prioritize lesion-related terminology over anatomical localization when processing highly specialized scientific language. This behavior helps explain the consistent confusion between digestive-system and neoplasm-related categories across models.

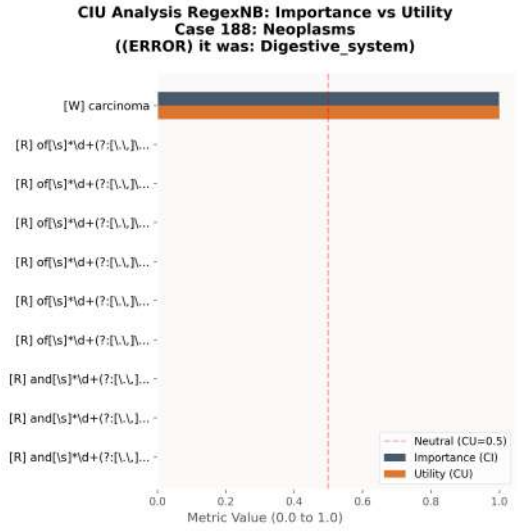

FIGURE 33. Local explainability analysis using the CIU method (MATC dataset - regexSetFit).

V. CONCLUSIONS

This work developed a hybrid architecture for biomedical text classification that balances predictive performance with the level of transparency required in biomedical and clinical domains. By integrating the CREGEX algorithm with machine learning and Transformer-based models such as BioBERT and SetFit, the proposed framework demonstrated that automatically generated logical rules can complement contextual semantic representations while preserving interpretability.

The experimental results obtained on the CWLC, the MIMIC-III Demo, and the MATC datasets confirm that combining explicit rule-based knowledge with statistical and contextual representations produces more robust feature spaces than isolated architectures. In particular, the incorporation of automatically generated RegExes improved the performance of traditional classifiers, especially in the NB architecture, where statistically significant improvements were observed across all evaluated datasets.

These findings directly address Research Question 1, demonstrating that automatically generated RegExes can improve the predictive performance of biomedical text classification models while simultaneously enriching the representation space with interpretable logical patterns.

The explainability analyses based on SHAP and CIU revealed substantial differences in how the evaluated architectures distribute feature importance. Traditional models such as regexNB and regexSVM exhibited stronger dependence on isolated high-frequency terms and explicit rule-based patterns, whereas Transformer-based architectures distributed attribution more broadly across contextual terms and automatically generated rules. Furthermore, the CIU analyses demonstrated that the proposed framework allows the contextual importance and utility of linguistic features to be analyzed at the individual prediction level, facilitating the identification of dominant lexical patterns, contextual ambiguities, and causes of classification errors.

These results provide a direct response to Research Question 2, confirming that hybrid architectures combining automatically generated RegExes with machine learning and Transformer-based models can produce interpretable predictions through complementary XAI methods such as SHAP and CIU.

The masking validation protocol was essential for evaluating the reliability of the attribution process and the dependence of each architecture on meaningful linguistic evidence. The experiments revealed that regexSVM exhibited the highest sensitivity to masking, particularly in the CWLC and MIMIC-III datasets, where removing highly attributed words and rules produced substantial reductions in predictive performance. In contrast, Transformer-based architectures such as regexBioBERT and regexSetFit preserved comparatively stable performance after masking, suggesting greater robustness through contextual semantic representations.

These observations directly answer Research Question 3, demonstrating that masking domain-relevant linguistic features significantly affects predictive performance, particularly in architectures that rely strongly on explicit lexical and rule-based evidence.

Overall, the proposed framework demonstrates that it is possible to integrate explainability, contextual semantic representations, and automatically generated logical rules into a unified biomedical text classification architecture. The results suggest that combining explicit and implicit knowledge sources improves interpretability while preserving competitive predictive performance across heterogeneous biomedical datasets.

As future work, the development of interactive XAI systems is proposed. Instead of relying exclusively on static attribution values, future architectures could incorporate interfaces that allow clinical experts to dynamically adjust or validate the importance assigned to automatically generated rules. This human-in-the-loop strategy could facilitate continuous refinement of the hybrid classifier through the integration of statistical evidence and expert biomedical knowledge.

ACKNOWLEDGMENT

This work was partly funded by UBB FAPEI FP2524107.

REFERENCES

[1] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[2] C. A. Flores, R. L. Figueroa, and J. E. Pezoa, "Active learning for biomedical text classification based on automatically generated regular expressions," *IEEE Access*, vol. 9, pp. 38767–38777, 2021.

[3] P. Báez, A. P. Arancibia, M. I. Chaparro, T. Bucarey, F. Núñez, and J. Dunstan, "Natural language processing for clinical text in spanish: The case of waiting lists in chile," *Revista Médica Clínica Las Condes*, vol. 33, no. 6, p. 576, 2022.

[4] C. A. Flores and R. Verschae, "A generic semi-supervised and active learning framework for biomedical text classification," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4445–4448, IEEE, 2022.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[7] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," in *NeurIPS 2022 Workshop on Efficient Natural Language Processing*, pp. 1–14, 2022.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[9] K. Främling, "Contextual importance and utility: a theoretical foundation," in *Australasian Joint Conference on Artificial Intelligence*, pp. 117–128, Springer, 2022.

[10] H. Sakai and S. S. Lam, "Large language models for health care text classification: Systematic review," *JMIR AI*, vol. 5, no. 1, p. e79202, 2026.

[11] R. Hara and T. Masada, "Evaluating fine-tuned modern llms beyond bert for medical text classification," in *IEICE Conferences Archives*, The Institute of Electronics, Information and Communication Engineers, 2024.

[12] L.-X. Zheng, S. Ma, Z.-X. Chen, and X.-Y. Luo, "Ensuring the correctness of regular expressions: A review," *International Journal of Automation and Computing*, vol. 18, no. 4, pp. 521–535, 2021.

[13] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, and P. Ge, "Regular expression based medical text classification using constructive heuristic approach," *IEEE Access*, vol. 7, pp. 147892–147904, 2019.

[14] C. Tu and M. Cui, "Learning regular expressions for interpretable medical text classification using a pool-based simulated annealing approach," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–7, IEEE, 2020.

[15] A. Sbei, K. ElBedoui, and W. Barhoumi, "Assessing the efficiency of transformer models with varying sizes for text classification: A study of rule-based annotation with distilbert and other transformers," *Vietnam Journal of Computer Science*, vol. 12, no. 03, pp. 301–328, 2025.

[16] X. Li, M. Cui, J. Li, R. Bai, Z. Lu, and U. Aickelin, "A hybrid medical text classification framework: Integrating attentive rule construction and neural network," *Neurocomputing*, vol. 443, pp. 345–355, 2021.

[17] D. D. Puri and G. Patnaik, "Regular expression-based text classification using msvm and machine learning techniques," in *Smart Data Intelligence: Proceedings of ICSMDI 2022*, pp. 199–210, Springer, 2022.

[18] N. Wagneur, O. Capitain, S. Supiot, F. Le Borgne, F. Bocquet, M. Campone, and T. Perennec, "Hybrid regex and natural language inference model as a zero-shot classifier for extracting data from medical reports," *JCO Clinical Cancer Informatics*, vol. 9, p. e2500130, 2025.

[19] C. A. Flores, R. L. Figueroa, J. E. Pezoa, and Q. Zeng-Treitler, "Cregex: A biomedical text classifier based on automatically generated regular expressions," *IEEE Access*, vol. 8, pp. 29270–29280, 2020.

[20] C. A. Flores and R. Verschae, "A hybrid method for clinical text classification based on confident predictions and regular expressions," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 064–069, IEEE, 2024.

[21] L. Campillos-Llanos, A. Valverde-Mateos, and A. Capllonch-Carrión, "Hybrid natural language processing tool for semantic annotation of medical texts in spanish," *BMC bioinformatics*, vol. 26, no. 1, p. 7, 2025.

[22] S. A. Lee, A. Wu, and J. N. Chiang, "Clinical modernbert: An efficient and long context encoder for biomedical text," *arXiv preprint arXiv:2504.03964*, vol. abs/2504.03964, 2025.

[23] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8990–9005, 2023.

[24] A. Dolk, H. Davidsen, H. Dalianis, and T. Vakili, "Evaluation of lime and shap in explaining automatic icd-10 classifications of swedish gastrointestinal discharge summaries," in *Scandinavian Conference on Health Informatics*, pp. 166–173, 2022.

[25] S. M. Dalhatu and M. A. A. Murad, "A model for enhancing pattern recognition in clinical narrative datasets through text-based feature selection and shap technique," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 4, pp. 2287–2296, 2024.

[26] H. Nie and X. Wu, "A dual-channel prediction-interpretation framework with pre-trained language models and shap explainability," *Journal of Computer and Communications*, vol. 13, no. 3, pp. 116–137, 2025.

[27] S. Fong, Z. Wang, G. C. Oliveira, X. Zhao, Y. Jiang, J. Liu, B.-L. Colton, S. W. Woods, M. Shenton, B. Nelson, *et al.*, "Chirpe: A step towards real-world clinical nlp with clinician-oriented model explanations," in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 646–658, 2026.

[28] K. Främling, "Explainable ai without interpretable model," *arXiv preprint arXiv:2009.13996*, vol. abs/2009.13996, 2020.

[29] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 740–770, 2021.

[30] K. Agrawal, R. El Shawi, and N. Ahmed, "Xai-eval: A framework for comparative evaluation of explanation methods in healthcare," *Digital Health*, vol. 11, p. 20552076251368045, 2025.

[31] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, *et al.*, "A review of explainable artificial intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109370, 2024.

[32] A. Malhi and K. Främling, "An evaluation of contextual importance and utility for outcome explanation of black-box predictions for medical datasets," in *World Conference on Explainable Artificial Intelligence*, pp. 544–557, Springer, 2023.

[33] A. A. Noor, A. Manzoor, M. D. Mazhar Qureshi, M. A. Qureshi, and W. Rashwan, "Unveiling explainable ai in healthcare: Current trends, challenges, and future directions," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 2, p. e70018, 2025.

[34] P. Báez, F. Villena, M. Rojas, M. Durán, and J. Dunstan, "The chilean waiting list corpus: a new resource for clinical named entity recognition in spanish," in *Proceedings of the 3rd clinical natural language processing workshop*, pp. 291–300, 2020.

[35] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[36] T. Schopf, D. Braun, and F. Matthes, "Evaluating unsupervised text classification: zero-shot and similarity-based approaches," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pp. 6–15, 2022.

[37] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific reports*, vol. 14, no. 1, p. 6086, 2024.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *CoRR*, vol. abs/1201.0490, 2012.

MAURICIO A. FUENZALIDA received the B.S. degree in computer science and informatics engineering. He is currently pursuing the M.S. degree in Computer science at Universidad del Bío-Bío, Chile. His research interests include text mining, machine learning and cibersecurity.

CHRISTOPHER A. FLORES received the professional degree in Biomedical Engineering from the Universidad de Concepción (UdeC), Concepción, Chile, in 2015. He also received the M.S. and Ph.D. degrees in Engineering Sciences, with a major in Electrical Engineering, from the same institution in 2017 and 2021, respectively. During his doctoral studies, he completed a research internship at the Center for Biomedical Informatics at George Washington University in Washington, D.C., USA. Subsequently, he completed postdoctoral fellowships in the Department of Electrical Engineering at UdeC and at the Institute of Engineering Sciences of the University of O'Higgins, Rancagua, Chile. Currently, he is a full-time professor in the Department of Electrical and Electronic Engineering at the University of the Bío-Bío. He collaborates in the first Ph.D. program in Artificial Intelligence in consortium with the four universities of the CRUCH Biobío-Ñuble. He has taught courses related to artificial intelligence, computing, and scientific programming. Additionally, he has contributed to various scientific publications and research projects in natural language processing, computer vision, machine learning, and biomedical informatics, with diverse applications in healthcare, education, and agriculture.

ROSA L. FIGUEROA received the B.Eng. degree from the University of Concepción, in 2004, and the Ph.D. degree in electrical engineering from the University of Concepción, in 2012. Her Ph.D. thesis explored different methods to obtain useful information from free text. She is currently a Faculty Member and a Researcher, in biomedical engineering degree part, with the Electrical Engineering Department, University of Concepción, and a Technical Board Member of the National Center on Health Information Systems. She has scientific publications in journals and conference proceedings. Her research interest is within the medical informatics area, mainly machine learning and text mining. She is currently working in research projects related to secondary use of medical data and text classification.

5. Conclusiones

Este trabajo abordó el desafío de la opacidad en la clasificación automática de textos biomédicos, proponiendo un marco de trabajo híbrido que integra la potencia de los modelos de aprendizaje automático y profundo con la transparencia de reglas lógicas basadas en expresiones regulares. A través del diseño de una arquitectura de doble canal y la implementación de protocolos de explicabilidad, se buscó transformar los sistemas de "caja negra" en herramientas técnicamente auditables, capaces de proporcionar una base de evidencia lingüística para la validación de expertos clínicos. Los hallazgos derivados de esta investigación permiten concluir lo siguiente:

5.1. Conclusiones Generales

5.1.1. Efectividad de la Arquitectura Híbrida

La investigación demuestra que la integración de un canal de reglas lógicas con modelos de aprendizaje automático y profundo mejora significativamente el desempeño en la clasificación de textos biomédicos; por ejemplo, al incorporar este canal al modelo Naïve Bayes (regexNB), la métrica F1-score aumenta de forma estadísticamente significativa en todos los conjuntos de datos, alcanzando hasta un 98.32 % en MIMIC-III. Se concluye que el framework propuesto logró un equilibrio óptimo entre el desempeño predictivo y la transparencia, superando las limitaciones de los modelos convencionales al proporcionar una justificación lógica verificable para cada clasificación realizada.

5.1.2. Impacto en Modelos Tradicionales vs Deep Learning

Se observó una dualidad en el beneficio de la hibridación, mientas que los modelos tradicionales experimentaron los incrementos más sustanciales en su desempeño predictivo al incorporar el canal de reglas, los modelos basados en Transformers mantuvieron su alta precisión pero adquirieron una capa de auditabilidad técnica esencial. Esto valida que la inyección de conocimiento explícito es una estrategia transversalmente efectiva, independiente de la complejidad estructural del clasificador base.

5.1.3. Validación de la Explicabilidad y Robustez

Mediante el uso de XAI y técnicas de enmascaramiento selectivo, se confirmó que las decisiones del modelo híbrido se fundamentan en patrones lingüísticos con significado clínico real. La degradación controlada del desempeño observada al ocultar términos de alta atribución demuestra que el sistema reduce la dependencia hacia sesgos estadísticos del conjunto de datos, constituyendo un avance crítico hacia la adopción de estas tecnologías en entornos de salud reales.

Referencias

[1] Alexander Dolk, Hjalmar Davidsen, Hercules Dalianis, and Thomas Vakili. Evaluation of lime and shap in explaining automatic icd-10 classifications of swedish gastrointestinal discharge summaries. In *Scandinavian Conference on Health Informatics*, pages 166–173, 2022.

[2] Christopher A Flores, Rosa L Figueroa, and Jorge E Pezoa. Active learning for biomedical text classification based on automatically generated regular expressions. *IEEE Access*, 9:38767–38777, 2021.

[3] Christopher A Flores, Rosa L Figueroa, Jorge E Pezoa, and Qing Zeng-Treitler. Cregex: A biomedical text classifier based on automatically generated regular expressions. *IEEE Access*, 8:29270–29280, 2020.

[4] Christopher A Flores and Rodrigo Verschae. A generic semi-supervised and active learning framework for biomedical text classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4445–4448. IEEE, 2022.

[5] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[7] Simon A Lee, Anthony Wu, and Jeffrey N Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*, abs/2504.03964, 2025.

[8] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41, 2022.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?.^{ex}plaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[11] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, 2023.

[12] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. In *NeurIPS 2022 Workshop on Efficient Natural Language Processing*, pages 1–14, 2022.