Preprint of:
Nawaz, S., Fournier-Viger, P, Wu, J. W.-M. (2024). SeqClin: Pattern-Based Analysis and Classification of Clinical Datasets. Proceedings of BIBM 2024, 6490-6497.

SeqClin: Pattern-Based Analysis and Classification of Clinical Datasets

M. Saqib Nawaz
Shenzhen University
Shenzhen, China
msaqibnawaz@szu.edu.cn

Philippe Fournier-Viger*
Shenzhen University
Shenzhen, China
philfv@szu.edu.cn

Jimmy Ming-Tai Wu
National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan
wmt@wmt35.idv.tw

***Abstract*—Accurate analysis and classification of clinical datasets are crucial for understanding disease patterns, identifying risk factors and devising targeted interventions that ultimately contribute towards effective healthcare systems and improved patient outcomes. However, existing analysis and classification methods often fall short of effectively capturing complex sequential relationships within patient data and have limited interpretability. To overcome these challenges, we introduce SeqClin, a novel approach that utilizes frequent pattern mining to obtain valuable sequential information from clinical datasets. SeqClin first transforms clinical datasets into an appropriate format. Then, it employs sequential pattern mining algorithms to find frequent sequential patterns as well as rules of patient features in the datasets. These identified feature patterns and their respective values are then used for classification/detection. The performance of SeqClin is evaluated on four clinical datasets, where six classification models and evaluation metrics are employed for a comprehensive assessment. The obtained results show that the proposed approach surpassed previous approaches, with the extracted patterns and rules providing valuable insights into the key patient features and their values in clinical datasets.**

***Index Terms*—Clinical data analysis, Frequent pattern mining, Sequential pattern discovery, Classification, Healthcare.**

I. Introduction

Clinical narratives and datasets play a crucial role in modern healthcare as they contain vast amounts of information, including patient medical histories, demographics, laboratory and biochemical results, imaging data, and treatment records [1]. Therefore, they offer not only useful insights into patients' health but also provide a rich context for clinical decision support systems (CDSS). These latter facilitate more informed and effective medical decision-making. By analyzing clinical datasets, healthcare professionals can make well-informed decisions and tailored treatments for individual patients. They can also discern patterns as well as trends and risk factors, all of which ultimately contribute to improving the overall quality of patient care [2].

The traditional method for clinical data analysis involves rule-based approaches [3], [4]. However, the development of such approaches is time-consuming and resource-intensive as they require a certain level of direct interaction with clinical experts to transform their often implicit knowledge into a comprehensive set of explicit rules. On the other hand, natural language processing (NLP), a subfield of artificial intelligence, has consistently proven its efficacy in extracting information from unstructured text. The rapid advancement of text analytics beyond foundational NLP capabilities has been significantly accelerated by the increased adoption of machine learning (ML) and deep learning (DL) methodologies. Over the past decade, they have been integrated into text analytics, enabling their application in diverse clinical care settings [5]–[8] to allow medical experts and clinicians to take proactive measures by predicting the diverse factors and reasons associated with a disease. In recent years, numerous studies [9]–[21] have focused on the development of computational approaches for the effective analysis and classification of diseases based on their clinical text data.

Some studies [9]–[13] focused on disease predictive modeling using ML and DL classifiers. The studies [14]–[21] identified significant features in the datasets such as features selected based on chaotic multi-verse optimization [14], genetic algorithm (GA)-based feature and instance selection [15], improved teacher-learner based optimization (ITLBO) [16], biogeography-based optimization [17], fruit fly optimization [18], boruta-based, ridge regression-based and random Fourier-based features [19], binning-based features [20], and manual feature-selection [21]. The feature extraction methods based on optimization techniques are computationally expensive, often resulting in long times for model learning. Moreover, most of the developed methods encounter challenges related to accuracy and interoperability, and their generalizability and scalability remain open questions, as their performance evaluation is done on a small number of datasets (typically only one). For clinical data, sophisticated yet simple and intuitive techniques that are based on frequent pattern mining (FPM) [22] are highly desirable because the clinical data depend on various factors/causes, and often exhibit atypical characteristics or properties.

FPM deals with finding interesting and important patterns within datasets. These patterns, which can manifest in diverse forms including rules, are interpretable due to their direct correlation and reference to attribute values present in the data. Two studies [23], [24] used the Apriori algorithm [**?**], a rule-based algorithm for FPM, to find feature patterns and rules in clinical datasets. Another study [25] proposed a constraint-based algorithm, based on CHARM [26], to find frequent

* corresponding author.

closed itemsets of symptoms, diagnosis and medication in a clinical dataset. However, Apriori and CHARM do not take into account the sequential arrangement of features. Sequential pattern mining (SPM) [27], on the other hand, offers efficient algorithms to analyze sequential data on the basis of frequent sequential rules and patterns. For instance, SPM has been used for the analysis and classification of biological data in genomic form [28]–[32]. As far as we know, there is a notable absence in the literature of any study that has incorporated the frequent sequential patterns of diverse patient features identified in clinical datasets into the classification/detection process.

In this paper, we introduce an SPM-based approach, called SeqClin (Sequential Clinical) for analyzing and classifying clinical datasets. SeqClin presents an integrated pipeline that starts with data transformation, transitioning clinical datasets into a proper format. Subsequently, the transformed datasets are processed via the algorithms for SPM to find recurrent feature sets and their corresponding values, along with the sequential relationships that exist among them. These relationships are represented in the form of sequential rules and patterns, providing valuable insights for the data. The frequent patterns found in the datasets are then utilized as features in the classification process. In total, six classifiers are used, and their overall performance is rigorously evaluated using a range of metrics.

The experimental findings demonstrate that utilizing SeqClin to identify frequent sequential patterns within clinical patient data and subsequently leveraging these discovered patterns results in enhanced classification performance. Notably, the random forest (RF) in SeqClin performed better than the other five classifiers and outperformed recent approaches for classifying clinical datasets. It was observed that frequent patterns and rules discovered in clinical datasets offer a deep understanding of the shared attributes and inherent characteristics of clinical text data. Thus, such patterns hold the potential to aid in the development of accurate, fast, and interpretable CDSS, thereby improving the accuracy and efficiency of clinical decision-making processes. Additionally, this research can also empower healthcare sectors to perform automated and insightful analyses, facilitating the extraction of essential information (key patient features) from clinical data where the order or sequencing of information is critical. This can facilitate the construction of vital knowledge bases, which can potentially lead to improvements in patient outcomes and enhance healthcare management.

The rest of this paper contains three sections: Section II provides the details about the SeqClin approach and the four clinical datasets. Section III presents and discusses experimental results and the comparison of SeqClin with recent approaches. Section IV concludes the paper with some remarks.

II. Methodology

The proposed SeqClin approach to analyze and classify/identify disease in clinical data (illustrated in Fig. 1) contains five parts: (1) Clinical datasets collection, (2) Datasets preprocessing, where features and their values are transformed into an appropriate format, (3) Feature extraction, in the forms of patterns and rules, using SPM algorithms, (4) Employing the extracted frequent sequential features and their values in the classification process, and (5) Benchmark evaluation to investigate the performance of the proposed approach. The next subsections provide the details for the first four parts.

A. Clinical Datasets

The performance and effectiveness of SeqClin are evaluated on four publicly accessible clinical datasets about various diseases. The first dataset is the Chronic Kidney Disease dataset[1], called CKD, which comprises 491 samples (or records) and 25 patient features. This dataset provides a set of information related to patients who are diagnosed with CKD, such as biochemical, clinical and demographic information. In this dataset, 435 and 56 patients are categorized as non-CKD and CKD (indicating their positive diagnosis of CKD) respectively. One feature, *StudyID*, was excluded from the analysis, as it contains the sequential IDs of patients and does not provide any important information. For this study, the CKD dataset contains 24 features, with 10 of these being numerical and the remaining 14 being categorical.

TABLE I
Statistical information about four clinical datasets

Dataset	Samples	Features (N/C)	Dependent Feature
CKD	491	24 (10/14)	EventCKD35 (56 Yes/435 No)
DSPP	349	10 (1/9)	Outcome(186 Positive/163 Negative)
CSD	95,984	17 (0/17)	CurrentStatus (87,951 LCP/8,033 PC)
HFP	918	12 (4/8)	HeartDisease (508 Yes/410 No)

N: Numerical, C: Categorical

The second dataset is the Disease Symptoms and Patient Profile (DSPP) dataset[2]. It contains 10 features, of which 9 are categorical and 1 is numerical. 186 patients in DSPP were identified as positive and 163 patients were identified as negative, after the diagnosis/assessment for the specific disease. DSPP comprises symptoms, demographics, and health indicators of patients.

The third dataset is the COVID-19 Surveillance dataset[3], referred to as CSD, available at the Center for Disease Control and Prevention CDC, USA[4]. It contains 95,984 samples after preprocessing, of which 87,951 belongs to laboratory-confirmed patients (LCP) and the remaining 8,033 belong to patients with probable case (PC). The CSD dataset originally had 19 categorical features. Two features, namely the state of residence (res_state) and county of residence (res_county), are not considered as two other features (state_fips_code and county_fips_code) provide the Federal Information Processing Standards (FIPS) code for states and counties. Thus, in this work, CSD includes 17 categorical features.

The fourth dataset is the Heart Failure Prediction (HFP) dataset[5], which comprises 918 records, from which 508

[1] figshare.com/articles/dataset/6711155?file=12242270

[2] kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset/data

[3] https:github.com/sarwanpasha/COVID_19_Clinical_Data_Analytics

[4] data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data

[5] kaggle.com/datasets/fedesoriano/heart-failure-prediction

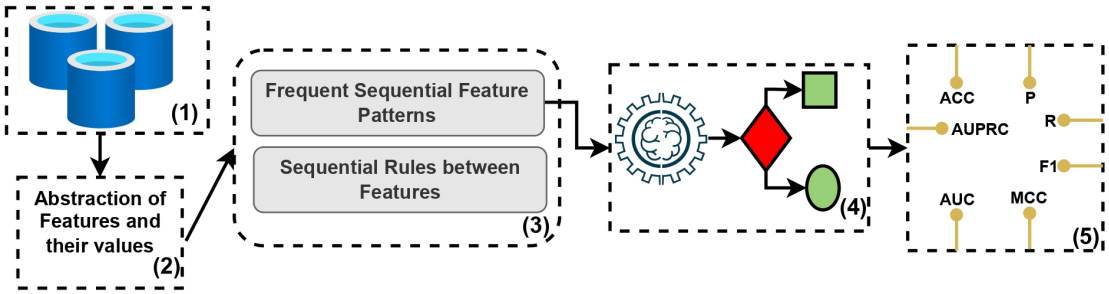


Fig. 1. The SeqClin approach for the analysis and classification of clinical datasets consists of five stages: (1) Datasets collection, (2) Dataset proprocessing and abstraction of features, (3) Discovery of (a) frequent sequential patient feature patterns and (b) rules among them in the datasets, (4) Performing classification by utilizing the identified frequent sequential patterns of patient features and their respective values to train various classifiers, and (5) Approach evaluation using various metrics.

records are patients diagnosed with heart disease and the remaining 410 without heart disease. This dataset contains 12 features, including 8 categorical and 4 numerical variables.

TABLE II
FEATURES IN THE CONSIDERED CLINICAL DATASETS

Dataset 1 (CKD)
(1) Gender, (2), Age, (3) AgeCategories, (4) HistoryDiabetes, (5) HistoryCHD, (6) HistoryVascular, (7) HistorySmoking, (8) HistoryHTN, (9) HistoryDLD, (10) HistoryObesity, (11) DLDmeds, (12) DMmeds, (13) HTNmeds, (14) ACEIARB, (15) Cholesterol, (16) Triglycerides, (17) HgbA1C, (18) Creatnine, (19) eGFR, (20) sBP, (21) dBP, (22) BMI, (23) TimeToEventMonths, (24) EventCKD35
Dataset 2 (DSPP)
(1) Disease, (2) Fever, (3) Cough, (4) Fatigue, (5) DifficultyBreathing, (6) Age, (7) Gender, (8) BloodPressure, (9) CholesterolLevel, (10) OutcomeVariable
Dataset 3 (CSD)
(1) CaseMonth, (2) StateFIPSCode, (3) CountyFIPSCode, (4) AgeGroup, (5) Sex, (6) Race, (7) Ethnicity, (8) CasePositiveSpecimenInterval, (9) CaseOnsetInterval, (10) Process, (11) Exposure, (12) CurrentStatus, (13) SymptomStatus, (14) Hosp, (15) ICU, (16) Death, (17) UnderlyingConditions
Dataset 4 (HFP)
(1) Age, (2) Sex, (3) ChestPainType, (4) RestingBP (5) Cholesterol, (6) FastingBS, (7) RestinECG, (8) MaxHR, (9) ExerciseAngina, (10) OldPeak, (11) STSlope, (12) HeartDisease

In the considered datasets, only CKD contains missing values for one feature in 21 samples. The other 3 datasets contain no missing values for features. Table I provides a statistical overview of the clinical datasets considered. Table II offers detailed feature information in sequential order within the dataset, with the dependent feature underlined.

B. Encoding

In the stage of data pre-processing, features within each dataset are transformed into a standardized *integer-based format* [33]. It was observed that various features across the clinical datasets share the same values. To eliminate any potential ambiguity, unique values of distinct features are encoded into distinct positive integers, which are henceforth referred to as feature values. In other words, within the transformed dataset, values that were originally identical for two distinct features in the original dataset are now represented by unique integers to facilitate differentiation.

Let $FV = \{FV_1, FV_2, ..., FV_m\}$ represent the list or set of patient features and their respective values in a dataset. Any subset, denoted as FVS ($FVS \subseteq FV$), is referred to as a *feature values list* or set. To efficiently explore the patterns search space in a dataset, a relation, denoted as $\prec$, is applied to the FV. This relation establishes a specific ordering on feature values, which ensures that SPM algorithms do not discover duplicate patterns [27].

A feature sequence $FS = \langle FVS_1, FVS_2, ..., FVS_n \rangle$ is a sequentially ordered list for which $FVS_i \subseteq FV$ and $1 \leq i \leq n$. A list containing more than one feature sequence makes a dataset that is referred to as a *feature values dataset* (FVD). For example, $FVD = \langle FS_1, FS_2, ..., FS_x \rangle$ is a feature values dataset containing x sequences with unique IDs ranging from 1 to x. A sample of the original HFP dataset is shown in Fig. 2(a), while its transformed version is shown in Fig. 2(b).

C. Discovering Frequent Sequential Patterns and Rules

The TKS [34] and ERMiner algorithms [35] are used to find frequent sequential patterns and rules associated with FV (patient features and their corresponding values). Both algorithms use certain measures to identify patterns such as the *support* (sup) and *confidence* ($conf$).

In a FVD, the number of sequences (Seq) that contain a particular Seq_p is referred to as the *support* of Seq_p, denoted by the symbol $sup(Seq_p)$:

$$sup(Seq_p) = |\{Seq | Seq_p \sqsubseteq Seq \wedge Seq \in FVD\}| \quad (1)$$

where $Seq_p \sqsubseteq Seq$ represents that Seq_p is present in Seq or in other words, Seq contains Seq_p. Formally, let there be two sequences Seq_p and Seq_q, defined as follows: $Seq_p = \langle p_1, p_2, ..., p_x \rangle$ and $Seq_q = \langle q_1, q_2, ..., q_y \rangle$. Then, Seq_p is considered a *subsequence* of Seq_q, if there exist an integers set $1 \leq n_1 < n_2 < ... < n_x \leq y$, such that $p_1 \subseteq q_{n1}, p_2 \subseteq q_{n2}, ..., p_x \subseteq q_{ny}$.

In a FVD, SPM algorithms do the complete enumeration of all the *frequent subsequences*. Formally, a sequence Seq is a *sequential pattern* (also known as a *frequent sequence*) if its support (sup) is greater than or equal to the minimum support (ms) threshold, that is set by the user ($sup(Seq) \geq ms$).

The main advantage of using TKS is that it offers a parameter k that users can specify to obtain a desired number of sequential patterns. On the other hand, traditional algorithms

for SPM necessitate the specification of a parameter *ms*. Using the parameter *k* of TKS offers the advantage of knowing the exact number of patterns that will be output prior to execution, thereby eliminating the need for multiple algorithm runs to achieve a specific number of patterns in a dataset. The process for generating candidates (subsequences) in TKS incorporates building a vertical database representation along with a depth-first search. These allow TKS to compute and count candidates efficiently without the necessity of a comprehensive scan of the dataset, thereby enhancing its performance when dealing with lengthy candidates or dense databases. Furthermore, TKS employs a range of advanced strategies to minimize the search space, such as the utilization of the PMAP (Precedence Map) data structure and an optimized join operation on a bit vector representation. For a comprehensive understanding of TKS, interested readers can find more details in [34]. A sample of frequent sequential patterns of patient features in both transformed and original formats is shown in Fig. 2(c) and Fig. 2(e), respectively.

On the contrary, sequential rules signify a correlation among two distinct items (patient features and their values here) sets, by considering not only an item's support (sup) but also its confidence ($conf$), also called the conditional probability. Thus, such rules provide a comprehensive understanding of how items are associated in data. Let r represents a sequential rule in FVD, having the form $r : M \rightarrow N$, that is an implication between two feature value sets $M, N \subseteq FV$ that are non-empty and disjoint. The sup and $conf$ of r are calculated as:

$$conf_{FVD}(r) = \frac{|\{Seq | r \sqsubseteq Seq \wedge Seq \in FVD\}|}{|\{Seq | M \sqsubseteq Seq \wedge Seq \in FVD\}|} \quad (2)$$

$$sup_{FVD}(r) = \frac{|\{Seq | r \sqsubseteq Seq \wedge Seq \in FVD\}|}{|FVD|} \quad (3)$$

A sequence $S_p = \langle p_1, p_2, .. , p_x \rangle$ contains M (shown as $M \sqsubseteq S_p$) if $M \subseteq \bigcup_{i=1}^{n} p_i$. Similarly, r is present in S_p (shown as $r \sqsubseteq S_p$) if a q (an integer) exists such that $1 \leq q < n$, $M \subseteq \bigcup_{i=1}^{q} p_i$ and $N \subseteq \cup_{i=q+1}^{n} p_i$. A rule r has meaning that items of M are typically followed by items of N.

Sequential rule mining within a dataset is the process of enumerating all the valid sequential rules present within the data. Formally, r is considered a *frequent sequential rule* if its support value is greater than or equal to a preset value ms ($sup_{FVCD}(r) \geq ms$) and r is a *valid sequential rule* if it occurs frequently and its confidence value is greater than or equal to a user-defined minimum confidence threshold mc ($conf_{FVCD}(r) \geq mc$). Note that both ms and $mc \in [0, 1]$.

The ERMiner algorithm [35], which stands for Equivalence class-based sequential Rule Miner, leverages a vertical database representation and exploits equivalence classes of rules sharing identical antecedents and consequents to comprehensively explore the rule search space. For further exploration of the search space, ERMiner searches for rules using two procedures named right and left merges. Additionally, a SDM (sparse data matrix) technique is utilized to prune the search space, enhancing ERMiner's efficiency compared to previous sequential rule mining algorithms. Unlike TKS, users need to specify two parameters *ms* and *mc* to obtain the desired number of rules. More details about ERMiner can be found in [35]. Fig. 2(d) and Fig. 2(f) show a sample of sequential rules of patient features in both transformed and original formats, respectively.

D. Classification

In this stage, the frequently occurring sequential patterns of patient features, with their respective values, are employed for the classification of diseases. The clinical datasets generally contain patient sequences (records) of various lengths. After carefully examining the clinical datasets, multiple consecutive occurrences of the features were observed.

From the perspective of patient-related data, such occurrences can be considered redundant as they do not provide significant insight or information for classification purposes. Therefore, during classification, repetitive patient features are treated individually. Moreover, many studies [14]–[20] used different feature selection methods to reduce the number of total features that are used in disease classification. High accuracy is achieved with reduced features as compared to using all the features in the datasets. Inspired by these observations, the proposed approach leverages the discovered frequent sequential patient patterns for the purpose of classification/detection in clinical datasets.

Six standard ML models are used for classification, including (1) Naive Bayes (NB), (2) Logistic Regression (LR), (3) Support Vector Machine (SVM), (4) k-nearest Neighbors (kNN), (5) Decision Tree (DT) and (6) Random Forest (RF). Their performance is evaluated by employing six metrics, including (1) Accuracy (ACC), which measures the overall correctness of predictions; (2) Recall (R), indicating the capability of a classifier to find all relevant instances; (3) Precision (P), reflecting the proportion of relevant instances among the retrieved ones; (4) F1 score (F1), indicates the harmonic mean of P and R; (5) Matthews Correlation Coefficient (MCC), a balanced metric that considers both true and false positives, as well as true and false negatives; and (6) Area Under the Curve (AUC), representing the classifier performance as its discrimination threshold is varied. The main reason for utilizing these metrics, which are defined next, is that they provide a comprehensive assessment of the performance of classifiers.

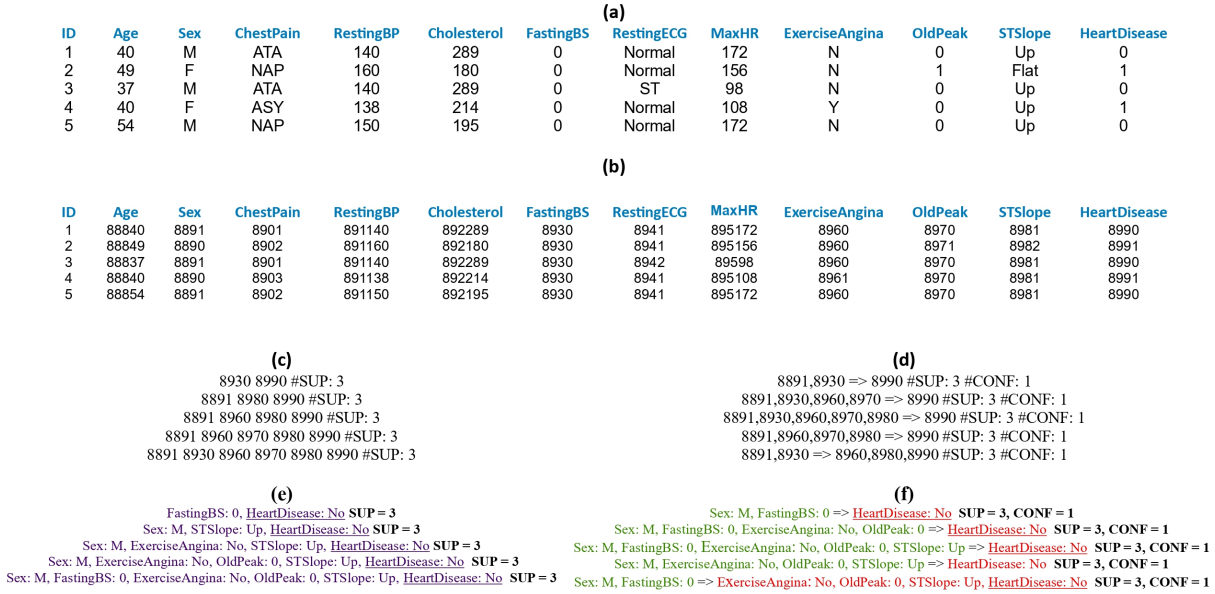
$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (5)$$

$$Precision(P) = \frac{TP}{TP + FP} \quad (6)$$

$$F - measure = 2 \times \frac{P \times R}{P + R} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$



(a)

ID	Age	Sex	ChestPain	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	OldPeak	STSlope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	140	289	0	ST	98	N	0	Up	0
4	40	F	ASY	138	214	0	Normal	108	Y	0	Up	1
5	54	M	NAP	150	195	0	Normal	172	N	0	Up	0

(b)

ID	Age	Sex	ChestPain	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	OldPeak	STSlope	HeartDisease
1	88840	8891	8901	891140	892289	8930	8941	895172	8960	8970	8981	8990
2	88849	8890	8902	891160	892180	8930	8941	895156	8960	8971	8982	8991
3	88837	8891	8901	891140	892289	8930	8942	89598	8960	8970	8981	8990
4	88840	8890	8903	891138	892214	8930	8941	895108	8961	8970	8981	8991
5	88854	8891	8902	891150	892195	8930	8941	895172	8960	8970	8981	8990

Fig. 2. The process of patient records transformation in a dataset and discovering frequent patterns as well as rules of patient features and their corresponding values. Some samples of the original dataset (a) are transformed into an abstracted dataset (b). Discovered patterns that occur frequently (c) and frequent rules (d) in the abstracted format. Discovered frequent sequential patterns (e) and rules (f) are presented in the original format. The dependent variable is underlined, green-colored features and their values in (f) represent antecedents, and red-colored features and their values represent consequent(s).

$$\text{AUC} = \int_0^1 \text{R}(d\text{FPR}) \tag{9}$$

where the terms TP, FP, TN, and FN stand for true positive, false positive, true negative and false negative, respectively. In equation 9, *dFPR* is for the derivative of $FPR = \frac{FP}{FP+TN}$. In this study, three validation strategies ((1) 5-fold cross-validation, (2) 10-fold cross-validation, and (3) 80:20 train-test split) are employed for robust evaluation of the classifiers' performance and for a comprehensive assessment of their generalizability and reliability. The next section discusses the fifth stage of the proposed methodology (evaluation).

III. Results and Discussion

A computing system with 16 GB of RAM and an Intel Core i5-11320H 3.20 GHz processor was utilized to conduct experiments. The SPMF library [36], developed in Java, was employed for pattern extraction within the abstracted clinical datasets. It offers many algorithms for the analysis and discovery of patterns, including TKS and ERMiner. Additionally, the open-source WEKA [37] software was utilized for the classifiers' training and testing.

A. Discovered Patterns and Rules

Table III presents several frequent sequential patterns of features that have been identified using TKS in the four clinical datasets, each characterized by varying lengths. These frequent sequential patterns offer invaluable insights into the prevalence and frequency of specific patient features and their corresponding values, providing a deeper understanding of patient characteristics and potential trends. For example, the first pattern indicates that approximately 84% of the patients with a *history of hypertension*, *who used medications* for this condition, have been diagnosed with CKD. Similarly, the patients with the following pattern *HistoryVascular: No*, *HistoryHTN: Yes*, *HistoryDLD: Yes*, *DLDmeds: No*, *HTNmeds: Yes*, *ACEIARB: Yes* have more than 60% chances of having *CKD*. HTN, DLD, CHD, D, and Meds stands for hypertension, dyslipidemia, coronary heart disease, diabetes, and medications respectively.

Table IV lists some of the discovered sequential rules by ERMiner in four datasets. These discovered rules reveal the intricate relationships among frequently occurring patient features and their respective values, offering a clearer picture of the underlying patterns and trends. Similar to frequent sequential patterns, discovered rules offer very useful information related to the interconnections between features and their values. Getting the desired number of rules in different datasets requires fine-tuning of the *ms* and *mc* parameters, along with two optional parameters for the maximum antecedent and consequent size.

The first rule in DSPP indicates that the patient features *Fatigue: Yes* and *DifficultyBreathing: No* are followed by *OutcomeVariable: Positive*. Similarly, the fourth rule in CSD suggests that a patient with *Race: White*, *Ethnicity: Non-Hespanic* and *CaseOnsetInterval: 0* was not submitted to ICU (*ICU: No*) and was a laboratory-confirmed case (*CurrentStatus: LCP*). With an occurrence frequency of 67.38%, this rule was prominently observed in the CSD. Moreover, the confidence level of 0.94 underscores the high correlation that exists between these patient features, indicating a high degree of certainty in the observed relationship. This suggests a 94% probability of a patient with the status of LCP when these specified conditions are met.

TABLE III
SEQUENTIAL PATTERNS DISCOVERED IN FOUR CLINICAL DATASETS

CKD: Yes	%
HistoryHTN: Yes, HTNmeds: Yes	83.92
HistoryHTN: Yes, HTNmeds: Yes, ACEIARB: Yes	76.78
HistoryHTN: Yes, HistoryDLD: Yes, HTNmeds: Yes, ACEIARB: Yes	69.64
HistoryHTN: Yes, HistoryDLD: Yes, DLDmeds: No, HTNmeds: Yes, ACEIARB: Yes	67.85
HistoryVascular: No, HistoryHTN: Yes, HistoryDLD: Yes, DLDmeds: No, HTNmeds: Yes, ACEIARB: Yes	60.71
CKD: No	%
HistoryCHD: No, HistoryVascular: No	89.19
HistoryCHD: No, HistoryVascular: No, HistoryDLD: No	77
HistoryDiabetes: No, HistoryCHD: No, HistoryVascular: No, DMmeds: No	57.24
HistoryDLD: No, HistoryHTN: Yes, HistoryDLD: Yes, DLDmeds: No, HTNmeds: Yes	34.71
HistoryCHD: No, HistoryDLD: No, HistoryHTN: Yes, HistoryDLD: Yes, DLDmeds: No, HTNmeds: Yes	30.34
DSPP: Positive	%
Fatigue: Yes, DifficultyBreathing: No	52.68
Fatigue: Yes, DifficultyBreathing: No, CholesterolLevel: Yes	36.02
Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male	26.34
Fever: No, Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male	20.43
Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male, BloodPressure: Yes, CholesterolLevel: Yes	13.44
DSPP: Negative	%
Fever: No, DifficultyBreathing: No	48.46
Fever: No, Fatigue: Yes, DifficultyBreathing: No	32.51
Fatigue: Yes, DifficultyBreathing: No, Gender: Male, CholesterolLevel: No	20.42
Fever: No, Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male	19.63
Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male, BloodPressure: No, CholesterolLevel: No	11.04
CSD: LCP	%
CaseOnsetInterval: 0, SymptomStatus: Symptomatic	98
CaseOnsetInterval: 0, SymptomStatus: Symptomatic, Death: No	90.81
SymptomStatus: Symptomatic, Hospital: No, ICU: No, Death: No	79.32
Race: Multiple/Other, CaseOnsetInterval: 0, Exposure: Yes, SymptomStatus: Symptomatic, ICU: No	64.32
Exposure: Yes, SymptomStatus: Symptomatic, Hospital: No, ICU: No, Death: No, UnderlyingConditions: Yes	57.79
CSD: PC	%
SymptomStatus: Symptomatic, Death: No	98.84
CaseOnsetInterval: 0, Hospital: No, Death: No	93.67
Ethnicity: Non-Hespanic, SymptomStatus: Symptomatic, Hospital: No, Death: No	88.41
StateFIPSCode: 39, CaseOnsetInterval: 0, SymptomStatus: Symptomatic, Hospital: No, Death: No	80.51
Race: Multiple/Other, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0, SymptomStatus: Symptomatic, Hospital: No, Death: No	77.92
HFP: Yes	%
Sex: Male, ChestPainType: ASY	69.48
ChestPainType: ASY, ExerciseAngina: Yes, STSlope: Flat	39.76
Sex: Male, ChestPainType: ASY, ExerciseAngina: Yes, STSlope: Flat	35.62
Sex: Male, ChestPainType: ASY, FastingBS: 1, ExerciseAngina: Yes, STSlope: Flat	25.78
ChestPainType: ASY, FastingBS: 1, RestingECG: Normal, ExerciseAngina: No, OldPeak: 0, STSlope: Flat	4.52
HFP: No	%
FastingBS: 1, ExerciseAngina: No	77.56
ExerciseAngina: No, OldPeak: 0, STSlope: Up	53.18
FastingBS: 1, ExerciseAngina: No, OldPeak: 0, STSlope: Up	49.26
Sex: Male, FastingBS: 1, RestingECG: Normal, ExerciseAngina: No, STSlope: Up	30.73
ChestPainType: ATA, FastingBS: 1, RestingECG: Normal, ExerciseAngina: No, OldPeak: 0, STSlope: Up	18.78

TABLE IV
SEQUENTIAL RULES DISCOVERED IN FOUR CLINICAL DATASETS

CKD			
Antecedents	**Consequents**	**%**	**Conf.**
HistoryVascular: No, HistoryHTN: Yes	EventCKD35: Yes	44.64	1
HistoryDLD: Yes, DLDMeds: Yes, HTNMeds: Yes	EventCKD35: Yes	33.92	1
HistoryHTN: Yes, HistoryDLD: Yes	DLDMeds: Yes, HTNMeds: Yes, EventCKD35: Yes	33.92	0.95
HistoryVascular: No, HistoryHTN: Yes, HTNMeds: Yes, ACEIARB: Yes	EventCKD35: Yes	37.5	1
HistoryDiabetes: Yes, HistoryVascular: No, HistorySmoking: No	HistoryHTN: Yes, HTNMeds: Yes, ACEIARB: Yes, EventCKD35: Yes	28.57	0.84
HistoryCHD: No, HistoryVascular: No	EventCKD35: No	41.49	1
HistoryVascular: No, HistorySmoking: No, DMMeds: No	EventCKD35: No	32.18	1
HistoryDiabetes: No, HistoryCHD: No	HistoryVascular: No, HistoryHTN: Yes, EventCKD35: No	17.93	0.61
Gender: Female, HistoryCHD: No, HistoryVascular: No, HistorySmoking: No	EventCKD35: No	29.66	1
HistoryCHD: No, HistoryVascular: No, HistorySmoking: No	HistoryDLD: Yes, DLDMeds: Yes, EventCKD35: No	23.44	0.62
DSPP			
Antecedents	**Consequents**	**%**	**Conf.**
Fatigue: Yes, DifficultyBreathing: No	OutcomeVariable: Positive	52.68	1
Fatigue: Yes, BloodPressure: Low, CholesterolLevel: 1	OutcomeVariable: Positive	36	1
Cough: No, Fatigue: Yes	DifficultyBreathing: No, Gender: Male, OutcomeVariable: Positive	26.34	0.60
Cough: Yes, DifficultyBreathing: No, BloodPressure: Low, CholesterolLevel: 1	OutcomeVariable: Positive	11.29	1
Fever: No, Cough: No, Fatigue: Yes, DifficultyBreathing: No, Gender: Male	OutcomeVariable: Positive	20.43	1
DifficultyBreathing: No, CholesterolLevel: 0	OutcomeVariable: Negative	44.17	1
Fever: No, Fatigue: Yes, DifficultyBreathing: No	OutcomeVariable: Negative	32.51	1
Cough: No, Fatigue: Yes	DifficultyBreathing: No, Gender: Male, OutcomeVariable: Negative	22.69	0.62
Fatigue: Yes, DifficultyBreathing: No, Gender: Male, CholesterolLevel: 0	OutcomeVariable: Negative	20.24	1
Fever: No, Cough: No, Fatigue: Yes, BloodPressure: Normal, CholesterolLevel: 0	OutcomeVariable: Negative	10.42	1
CSD			
Antecedents	**Consequents**	**%**	**Conf.**
Race: White, CaseOnsetInterval: 0	CurrentStatus: LCP	85.67	1
Race: White, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0	CurrentStatus: LCP	71.03	1
CaseOnsetInterval: 0, Exposure: Yes	ICU: No, CurrentStatus: LCP	72.71	0.94
Race: White, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0	ICU: No, CurrentStatus: LCP	67.38	0.94
Sex: Female, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0, Exposure: Yes	CurrentStatus: LCP	40.66	1
Ethnicity: Non-Hespanic, CaseOnsetInterval: 0	CurrentStatus: PC	90.094	1
Ethnicity: Non-Hespanic, CaseOnsetInterval: 0, Exposure: Yes	CurrentStatus: PC	71.24	1
StateFIPSCode: 39, Race: White, Ethnicity: Non-Hespanic	CaseOnsetInterval: 0, CurrentStatus: PC	70.14	0.99
Race: White, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0	Exposure: Yes, CurrentStatus: PC	64.33	0.78
StateFIPSCode: 39, Race: White, Ethnicity: Non-Hespanic, CaseOnsetInterval: 0	CurrentStatus: PC	70.14	1
HFD			
Antecedents	**Consequents**	**%**	**Conf.**
Sex: Male, ChestPainType: ASY	HeartDisease: Yes	69.48	1
Sex: Male, ChestPainType: ASY, STSlope: Flat	HeartDisease: Yes	50.98	1
FastingBS: 0, ExerciseAngina: Yes	STSlope: Flat, HeartDisease: Yes	35.43	0.8
Sex: Male, ChestPainType: ASY, RestingECG: Normal, STSlope: Flat	HeartDisease: Yes	30.51	1
ChestPainType: ASY, FastingBS: 0, RestingECG: Normal, ExerciseAngina: Yes, STSlope: Flat	HeartDisease: Yes	18.11	1
ExerciseAngina: No, STSlope: Up	HeartDisease: No	70.97	1
FastingBS: 0, ExerciseAngina: No, STSlope: Up	HeartDisease: No	64.14	1
Sex: Male, FastingBS: 0	ExerciseAngina: No, HeartDisease: No	48.53	0.85
RestingECG: Normal, ExerciseAngina: No, OldPeak: 0, STSlope: Up	HeartDisease: No	38.53	1
Sex: Male, RestingECG: Normal, ExerciseAngina: No, STSlope: Up	HeartDisease: No	33.41	1

TKS and ERMiner algorithms enable the identification of not only frequently occurring features and their interrelationships but also the significance of specific feature values that could potentially impact patient analysis in clinical datasets. Note that the frequent sequential patterns and rules found in clinical datasets can be interpreted or considered as their descriptors or features. Such features and rules can be used in the classification process instead of providing all the available patient features within a dataset. As demonstrated in III-B, adopting this strategy can not only streamline the classification procedure but potentially enhance its accuracy and efficiency.

B. Classification Results

Prior to classification, the identified frequent sequential patterns within four datasets are subjected to preprocessing to guarantee that each pattern's length meets a minimum threshold of four. The standard (default) hyperparameters of classifiers in WEKA are utilized to streamline the model training process and establish a benchmark for performance comparison. The classification results for the six models, when trained and tested using three validation strategies, are presented in Table V.

Two classifiers (DT and RF) outperformed the other four classifiers (NB, LR, SVN and kNN) across all four datasets when evaluated using the three validation strategies. The paired t-test further validated that RF and DT significantly outperformed the other four classifiers. RF exhibited slightly better performance when using an 80:20 splitting strategy compared to 5-fold and 10-fold cross-validation. Conversely, DT performed better on 10-fold cross-validation. NB achieved better results on datasets when using 10-fold cross-validation. Meanwhile, LR, SVM and kNN demonstrated similar performance across all four datasets, with negligible differences

TABLE V
OBTAINED ACCURACY OF SIX CLASSIFIERS ON FOUR CLINICAL DATASETS

Validation	Dataset	NB	LR	SVM	kNN	DT	RF
5-fold	CKD	0.549	0.606	0.599	0.566	0.996	**1**
	DSPP	0.495	0.51	0.502	0.571	0.995	**0.997**
	CSD	0.652	0.655	0.657	0.77	0.994	**0.998**
	HFP	0.53	0.615	0.606	0.581	**0.998**	0.997
	Average	0.556	0.596	0.591	0.622	0.995	**0.998**
Validation	**Dataset**	**NB**	**LR**	**SVM**	**kNN**	**DT**	**RF**
10-fold	CKD	0.554	0.596	0.596	0.566	0.998	**1**
	DSPP	0.998	0.509	0.518	0.578	0.994	**0.999**
	CSD	0.967	0.65	0.657	0.793	0.994	**0.998**
	HFP	0.531	0.614	0.611	0.553	**0.998**	0.996
	Average	0.762	0.592	0.595	0.622	0.996	**0.998**
Validation	**Dataset**	**NB**	**LR**	**SVM**	**kNN**	**DT**	**RF**
80:20	CKD	0.61	0.63	0.59	0.605	**1**	**1**
	DSPP	0.45	0.505	0.50	0.53	0.995	**1**
	CSD	0.725	0.73	0.73	0.77	0.99	**1**
	HFP	0.56	0.605	0.58	0.585	**0.99**	**0.99**
	Average	0.586	0.617	0.6	0.622	0.993	**0.999**

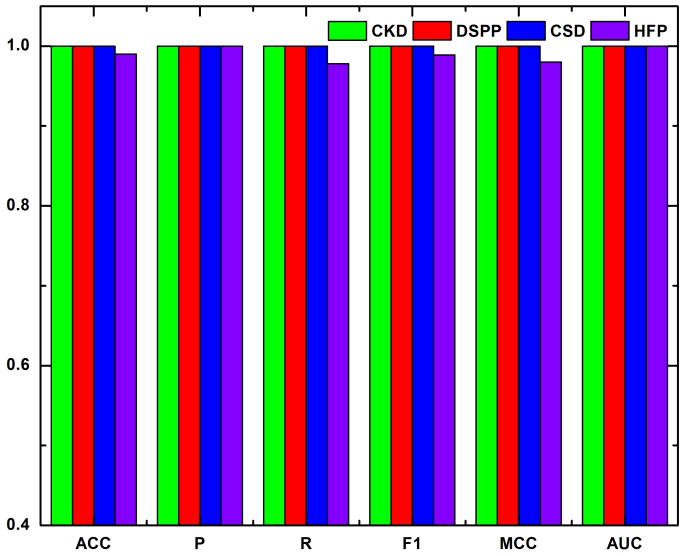


Fig. 3. RF results on four clinical datasets

observed between the three validation strategies. All of the classifiers performed efficiently and completed their execution within seconds. Based on the average accuracy across all four datasets and three validation strategies, the classifiers have been ranked as (1) RF (99.8%), (2) DT (99.4%), (3) NB (63.4%), (4) kNN (62.2%), (5) LR (60.1%), (6) SVM (59.5%).

The comprehensive results of the RF classifier, utilizing an 80:20 splitting ratio, across four clinical datasets are presented in Fig. 3. The main finding of this work is that instead of using all the patient features, frequent sequential patterns discovered in clinical datasets can be effectively employed in the identification/detection process. The TKS patterns used in the classification process consist of no more than 9 features.

The proposed approach is compared in Table VI with recent methods published in 2020-2024. While most of the methods [10]–[13], [16]–[18], [20], [21] utilized a single dataset for evaluation, other [9], [15], [19] utilized multiple datasets. The study [15] employed the largest number of seven clinical datasets. DNN, XGB, CL and CDD stand for Deep Neural Network, XGBoost, CNN+LSTM and Cardiovascular Disease Dataset, respectively. A Multi-Kernel SVM (MKSVM) was used in [18]. Among the three feature selection methods considered, [19] achieved better performance when RF was applied to Boruta-based features.

TABLE VI
COMPARISON OF SEQCLIN WITH RECENT APPROACHES

Model	Dataset	ACC	P	R	F1	MCC	AUC
DNN [9]	HFP	0.981	0.973	0.986	0.98	0.96	–
RF [10]	HFP	0.88	0.84	0.97	0.90	–	–
XGB [11]	CKD	0.932	–	0.918	0.931	–	0.968
CL [12]	CSD	0.923	0.923	0.936	0.93	–	0.90
RF [13]	HFP	0.956	0.552	0.976	–	–	–
RF [15]	HFP	0.968	0.944	1	0.933	–	–
CNN [16]	CKD	0.952	–	0.968	0.96	–	–
XGB [17]	CKD	0.983	1	0.973	0.986	–	–
SVM [18]	CKD	0.985	1	0.976	–	–	–
RF [19]	CSD	0.95	0.94	0.95	0.94	–	0.72
MLP [20]	CDD	0.872	0.887	0.848	0.867	–	0.95
XGB [21]	CSD	0.96	0.93	1	0.97	–	0.974
SeqClin(RF)	CKD	**1**	**1**	**1**	**1**	**1**	**1**
	CSD	**1**	**1**	**1**	**1**	**1**	**1**
	HFP	**0.99**	**1**	**0.987**	**0.989**	**0.98**	**1**

Notably, the proposed approach outperformed recent methods across all six evaluation metrics. On HFP and CKD datasets, the RF in the proposed approach achieved an improvement of approximately 1% over DNN [9] and 1.73% on XGB [17]. respectively. On the CSD dataset, RF achieved an improvement of approximately 5% over RF [19]. DNN [9] results for Stalog HFP and RF [19] results for CSD in Table VI is for 5-fold cross-validation. The study [17] first divided the CKD dataset into four subsets based on missing values and feature selection method. XGB achieved a higher accuracy of 98.33% on 'Set 4' (containing 13 features out of 24) for 70:30 split when BBO was used for feature selection. Note that XGB [17] achieved the highest accuracy of 99.16% on 'Set 2' when all the features were used in classification. Note that the study [21] used a private CSD dataset and the CSD in [12] is different from the one we used in this work.

In summary, the use of SPM effectively captures complex sequential relationships in clinical data. SeqClin is computationally efficient and scalable, completing execution within minutes. The *integer-based* data transformation simplifies processing while maintaining semantic information. However, SeqClin requires parameter tuning for desired pattern extraction and its computational complexity increases with the number of patterns. In its current form, Seqclin does not explicitly address class imbalance or missing value issues in clinical datasets.

IV. CONCLUSION

This paper proposed a novel approach (called SeqClin) based on pattern mining for analyzing and classifying clinical datasets. The proposed approach was evaluated on four diverse datasets that were initially transformed into integer-based format. After that, algorithms for SPM were employed on the transformed datasets to identify frequent patient features and their corresponding values in the form of patterns and rules. Lastly, the discovered frequent sequential patterns were subsequently utilized in the classification task. The performance of the six classifiers for the classification tasks was investigated by using six evaluation metrics. The obtained

results indicated that RF outperformed the other five in the classification. Additionally, the proposed approach outperformed recent methods. This study acknowledges several limitations: (1) The retrospective and static nature of the clinical datasets posed challenges in ensuring standardization of features, as the majority of features vary across datasets without a specified value range. (2) The online origin of the datasets may lead to potential information collection bias. (3) The discovered frequent sequential patterns and rules require further validation, verification, and confirmation by medical experts and clinicians.

For future work, one key area is employing SeqClin on more clinical datasets, particularly dynamic datasets, to identify which discovered frequent patterns significantly contribute to the classification process. Additionally, incorporating oversampling and imputation techniques could improve data balancing and address missing values. Another direction is the development of classifiers based on the discovered sequential rules. Lastly, exploring contrasting frequent patterns [38] and employing them in the classification process could provide additional insights and improvements.

REFERENCES

[1] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead and K. B. Johnson, "Data from clinical notes: A perspective on the tension between structure and flexible documentation", J. Am. Med. Inform. Assoc, vol. 18, no, 2, pp. 181-186, 2011.

[2] I. Spasic and G. Nenadic, "Clinical text data in machine learning: Systematic review", JMIR Med. Inform, vol. 8, no. 3, e17984, 2020.

[3] I. Spasic, J. Livsey, J. A. Keane and G. Nenadic, "Text mining of cancer-related information: Review of current status and future directions", Int. J. Med. Inform, vol. 83, no. 9, pp. 605-623, 2014.

[4] M. A. Roberts and B. H. Abery, "A person-centered approach to home and community-based services outcome measurement", Front. Rehabil. Sci, Vol. 4, 2023.

[5] S. H. Lee, "Natural language generation for electronic health records", npj Digital Med, vol. 1, 63, 2018.

[6] C. Comito, D. Falcone and A. Forestiero, "Current trends and practices in smart health monitoring and clinical decision support", in Proc. of BIBM, 2020, pp. 2577-2584.

[7] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence", Nat. Med, vol. 25, pp. 44–56, 2019.

[8] S. Bharati, M. R. H. Mondal and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?", IEEE Trans. Artif. Intell, vol. 5, no. 4, pp. 1429-1442, 2024.

[9] S. I. Ayon, Md. M. Islam and Md. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques", IETE J. Res, vol. 68, no. 4, pp. 2488–2507, 2020.

[10] N. S. M. Huang, Z. Ibrahim and N. M. Diah, "Machine learning techniques for early heart failure prediction", Malaya. J. Comput, vol. 6, no. 2, pp. 872 - 884, 2021.

[11] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases", Sci. Rep, vol. 14, 3687, 2024.

[12] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection", Chaos Soliton. Fract, vol. 140, 110120, 2020,

[13] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis", Health Technol, vol. 11, pp. 87–97, 2021.

[14] S. Sreejith, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection" Comput. Bio. Med, vol. 126, 103991, 2020.

[15] V. R. E Christo, H. K. Nehemiah, J. Brighty, and A. Kannan, "Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest", IETE J. Res, vol. 68, no. 4, pp. 2508–2521, 2020.

[16] M. Manonmani and S. Balakrishnan, "Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset", Procedia Comput. Sci, vol. 171, pp. 1660-1669, 2020.

[17] M. J. Raihan, Md. A. M. Khan, S. H. Kee and A. A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP", Sci. Rep, vol. 13, 6263, 2023.

[18] J. J. Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm", Int. J. Imaging Syst. Technol, vol. 30, pp. 660–673, 2020.

[19] S. Ali, Y. Zhou and M. Patterson, "Efficient analysis of COVID-19 clinical data using machine learning models", Med. Biol. Eng. Comput, vol. 60, pp. 1881–1896, 2022.

[20] C. M. Bhatt, P. Patel, T. Ghetia and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques", Algorithms, vol. 16, no. 88, 2023.

[21] MN. Mohebbi, M. Tutunchian, M. M. Kargari, and A. B. Kharazmy, "Comparing supervised machine learning models for COVID-19 patient detection using a combination of clinical and laboratory dataset" Int. J. Web Res, vol. 5, no. 1, pp. 8-18, 2022.

[22] C. C. Aggarwal, M. A. Bhuiyan and M. A. Hasan, "Frequent pattern mining algorithms: A survey", in C. Agarwal and J. Han, Eds, Frequent Pattern Mining, Springer, 2014, pp. 19-61.

[23] A. Alaiad, H. Najadat, B. Mohsen and K. Balhaf, "Classification and association rule mining technique for predicting chronic kidney disease", J. Infor. Know. Manag, vol. 19, no. 1, 2040015, 2020.

[24] M. Tandan, Y. Acharya, S. Pokharel and M. Timilsina, "Discovering symptom patterns of COVID-19 patients using association rule mining", Comput. Bio. Med, vol. 131, 104249, 2021.

[25] J. Yu, L. Zhang, N. Xu, L. Fa and K. Yang, "Application of constraint-based frequent closed itemsets mining in TCM clinical data analysis", in Proc. of BIBM, 2023. pp. 4689-4696

[26] M. J. Zaki and C. J. Hsiao, "CHARM: An efficient algorithm for Closed itemsets Mining", in Proc. of SDM, 2002, pp. 457-473.

[27] P. Fournier-Viger, J. C. W. Lin, R. U. Kiran, Y. S. Koh and R Thomas, "A survey of sequential pattern mining", Data Sci. Patt. Recog, vol. 1, no. 1, pp. 55-74, 2017.

[28] M. S. Nawaz, P. Fournier-Viger, S. Nawaz, W. Gan and Yulin He, "FSP4HSP: Frequent sequential patterns for the improved classification of heat shock proteins, their families, and sub-types", Int. J. Biol. Macromol, vol. 277 (Part 1), 134147, 2024.

[29] M. S. Nawaz, P. Fournier-Viger, S. Nawaz, H. Zhu and Unil Yun, "SPM4GAC: SPM based approach for genome analysis and classification of macromolecules", Int. J. Biol. Macromol, vol. 266 (Part 2), 130984, 2024.

[30] M. S. Nawaz, P. Fournier-Viger, Y. He and Q. Zhang, "PSAC-PDB: Analysis and classification of protein structures", Comput. Bio. Med, volume 158, 106814, 2023,

[31] M. S. Nawaz, P. Fournier-Viger, A. Shojaee and H. Fujita, "Using artificial intelligence techniques for COVID-19 genome analysis", Appl. Intell, vol. 51, pp. 3086–3103, 2021.

[32] M. S. Nawaz, P. Fournier-Viger, M. Aslam, W. Li and X. Niu, "Using alignment-free and pattern mining methods for SARS-CoV-2 genome analysis", Appl. Intell, vol. 53, pp. 21920–21943, 2023.

[33] M. S. Nawaz, M. Z. Nawaz, P. Fournier-Viger, J. M. Luna, Analysis and classification of employee attrition and absenteeism in industry: A sequential pattern mining-based methodology, Comput. Ind, vol. 159–160, 2024.

[34] P. Fournier-Viger, A. Gomariz, T. Gueniche, E. Mwamikazi and R. Thomas, "TKS: Efficient Mining of Top-K Sequential Patterns", in Proc. of ADMA, LNCS, vol 8346. Springer, 2013, pp 109–120.

[35] P. Fournier-Viger, T. Gueniche, S. Zida and V. S. Tseng, "ERMiner: Sequential rule mining using equivalence classes", in Proc. of IDA. LNCS, vol 8819, Springer, 2014, pp. 108–119.

[36] P. Fournier-Viger, J. C. W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng and H. T. Lam, "The SPMF open-source data mining library version 2", in Proc. of ECML PKDD, LNCS, vol. 9853, 2016, pp. 36-40.

[37] E. Frank, M. A. Hall and I. H. Witten, Data Mining: Practical Machine Learning Tools and Techniques, 4th Ed, Morgan Kaufmann, 2016.

[38] S. Ventura and J. M. Luna, Supervised Descriptive Pattern Mining, Springer, 2018.