International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228646

Rakel Model For Multi Class Label Classification Using Ensemble Neural PCA On Healthcare Event Log

Smt. S. Yamuna Rani[1], Dr. Sumagna Patnaik[2]

[1]Department of Computer Science, Govt.Degree College for Women, Gajwel, Telangana, India

[2]Professor of Computer Science, J. B. Institute of Engineering and Technology, Hyderabad, Telangana, India

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022

Page Number : 334-342

Article History
Accepted: 12 Nov 2022
Published: 29 Nov 2022

ABSTRACT

Process change over time is of particular issue in the field of healthcare, as healthcare practices emerge and change in response to the individual needs of patients. We propose a systematic procedure to study the change in process in time, which is appropriate for the complex field of healthcare. Our approach is based on qualitative process comparison that is based on 3 levels: A broad viewpoint (process model) and a mid-level perspective (trace) and a fine-grained, detailed (activity). Our goal was to identify the changes, and understand the process's evolution. We demonstrate this approach by through a case study of tumor pathways within Leeds where we observed evidence of change points at various levels. This paper will expand our investigation by using redundancy strategies employing Neural PCA. We labeling the labels in order to determine and analyzing the miners utilized in process discovery. We also provide an in-depth analysis of the process of research at the trace and activity levels using group classifiers. Through our study we demonstrate that this approach is qualitative and can provide a valuable understanding of changes in process in time. Analyzing change on three levels will provide evidence for the process's evolution when different perspectives agree and contradictory evidence may result in a discussion with experts in the field. This approach is useful to those who are dealing with complex processes that undergo changes in time.

Keywords : Process Mining, PCA, Neural, Ensemble, treatment, Health care, Cancer Data.

I. INTRODUCTION

Projects in research on process mining just like other projects in data analytics, use data that has been that is collected over a period of months or even years. These projects are typically initiated by making the assumption that no modifications to the processes over the time of the study. However, there is the possibility that there were changes in both the process itself and in the data produced by the process. For healthcare facilities this is a major problem because care patterns change and develop according to the individual patient's demands, treatment procedures as well as Electronic Healthcare Record

(EHR) system modifications, in addition to other factors. These changes are triggered by complicated interactions between people as well as processes, technologies and the changing structure of organizations. It is essential to understand and analyze the process's changes as time passes in the field of healthcare process mining, to ensure that the intended changes can be tracked and unintended modifications can be investigated and recognized.

Process mining refers to a process-based method of data science that makes use of event logs for identifying and analyzing business processes models. A log of events is an account of activities that are time stamped through the system. Process mining is used to healthcare processes to ensure improving patient safety, quality improvement and resource optimization in the healthcare setting.

As a broader group of illnesses cancer is a complex disease and can be affecting any part in the body. There are at most 65 types of cancer that are recognized. The most prevalent cancer among women, affecting around 12 percent of women across the world. In the UK the breast cancer rate is among the four most prevalent cancers, including lung cancer, prostate cancer and colorectal cancer. The diagnosis of breast cancer is made through physical examination, mammogram and ultrasound scans, MRI, blood chemistry studies and biopsies of the affected part within the breast. Surgery is the main treatment option, and it can be followed by radiation therapy or chemotherapy therapy or both. The chemotherapy course is typically completed in six cycles in which each cycle is given 21 days following the preceding one. Certain patients may not be able an entire cycle of chemotherapy because of negative events, such as the need for emergency admission or neutropenia.

Concept drift is used by the machine learning community to describe the evolution of processes as they progress over time. The Process Mining community has used this term to describe the changing character of the procedure, or the details that are recorded on the procedure. There is an increasing amount of research that explores new ways of analyzing concepts drift [2,3,4]. There are three main challenges to overcome when it comes to conceptual drift. (1) The detection of change points (2) Change localization and characterization as well as (3) The evolution of a change process. The process of detecting change points is all about finding out if a process has changed while localization is about determining when and in which the change took place. Change characterization is a method to comprehend the characteristics of a change and identify the key elements of a change in a process. Change process evolution seeks to determine the progression of a process through time.

Change analysis of processes is usually accomplished by constructing models of processes of different time frames in the vast dataset, and then comparing them to detect changes in the process. This method is a kin to the process of comparison that is widely used to check conformance when the reference model is contrasted with the event log that records the actual execution of the process [5]. A process comparison method for analysis of process changes proposed by Partington et al. [6] also included the definition of the points of comparison based on different metrics. However Partington and al.'s approach was designed to evaluate processes across different hospitals, and it isn't specifically relevant to analysis of process change. Additionally Partington and al.'s method required clinical knowledge to choose particular clinical metrics that differ between various clinical domains. Bolt et al. [7] A different process comparison method was proposed using the different the frequency of activity and the percentages within the logs. This method allowed for a thorough comparison of every activity in two logs however, it did not allow direct comparison between processes. Both work were carried out by Partington and colleagues. and Bolt et

al. They aren't directly connected to analysis of process changes. However, the basic concepts of comparison are useful to analyze changes in processes over time. Although each of these methods offer a glimpse into the changes in processes however, there is no one universal method that can benefit from the benefits and perspectives of each.

In this paper, we describe an exploratory study in which we identify and analyse the changes in time within complex longitudinal data on healthcare. The goal was to use an approach at multiple levels to recognize the process, identify it and localise changes. This calls for a dimension reduction or reduction of data method employing Neural PCA to classify data using the multi classifier, and finally an Rakel Ensmeble model to aid in data mining.

II. RELATED WORK

2.1. Process Discovery Algorithms

There are just a few reviews of the processing of data and mining within the healthcare field. We've found a handful of reviews that focus on the application of data mining in different medical areas [17-23] and when it comes to process mining in medical care, there's an extremely brief and precise review of research studies that focus on the clinical path [24, 25]. However, there's no comprehensive study that assembles the specifics, descriptions and contexts of each case study that has examined how process mining was employed in the field of healthcare.

Process mining is the use of techniques that aim to extract useful information from the data that processes generate as they carry out. It serves to act as an interfacing between the area of science and process (which includes areas such as operations management and business processes research) as well as data science (which includes areas like predictive analytics, data mining and so on) and provides ways of analyzing processes using data [8]. Process mining isn't an exclusive approach to domains, i.e, process mining methods can be applied to any area in which processes are utilized and relevant data are readily accessible. Healthcare as the subject of this article, is one particular field where processing mining's use is growing.

Process mining in the healthcare sector is the context of the whole. Process mining encompasses the various processes which can be explained with an overall process model such as an image which represents the different steps that are involved within the course of the operation, and the various routes that the process may follow [22,23]. The steps in the process can be represented in a variety of ways e.g. by using flow diagrams [23] and Business Process Modeling notation (BPMN).

An enormous effort is evident in the production of scientific research which is connected to Health Technology Assessment for healthcare management. There's a lot to study in relation to the most recent methods of healthcare and this calls for the development of appropriate methodologies and processes that are tailored to the specific context of each country and the specific circumstances of each. This includes designing and implementing cutting-edge Health Technology Assessment techniques to aid in the healthcare decision-making process more efficient. [2,5].

A few authors have highlighted the difficulties in integrating findings of Health Technology Assessment studies undertaken in various countries due to potential for reproducibility as well as replicability i.e. there is a variation in the efficacy of the various methods, and the use of resources within the healthcare system, as well as the impact of epidemiological issues such as. To address this issue, the authors recommend a number of research studies that are based on data from clinical studies as well as information about the use of the particular region. [6-8].

III. DATASET & PREPROCESSING METHODS

The patient Pathways Manager (PPM) System incorporates data from various systems in the e Leeds Teaching Hospitals NHS Trust (LTHT) which includes admissions of patients, treatments ((chemotherapy or surgery), radiotherapy). Pathology and investigations, Multidisciplinary Team (MDT) meetings, consultations, as well as outpatients.

The PPM database holds clinical data regarding all patients in the hospital as well as cancer patients. We were granted access to the PPM database by using the IRAS application that provides the direct connection to a safe SQL database running on the virtual machine. The data was screened as well as cleaned and consolidated prior to being approved to accessibility by researchers. The PPM database is comprised of the clinical information from more than three million people including more than 270,000 are diagnosed with cancer at the very least.

The log of access to clinical users is stored in the PPM Splunk. The PPM Splunk is a web-based application management system that records real-time user login to the PPM system. This helps in analyzing the use of specific functions in the system. Each time a user accesses information in PPM's EHR system it automatically records activities within the PPM Splunk. In this study the access log for healthcare users log was primarily focused on that GPTab log to show the functions that are related to treatment for cancer. GPTab is a feature that allows doctors to search for patient records within the GP system, in order to assist with clinical decisions regarding patient treatment.

Statistics employ sampling techniques that are extremely helpful to learn about the behavior of the people under study. Random normal distributions of classes is the most common method for evaluation of any group that is under investigation. The conventional approach has been thought to be less laborious in binomial issues or any prognostic data that only includes positive and negative classes of health prediction.

The SEER data collection is a huge repository, and understanding the behaviour that the SEER cancer dataset exhibits is a problem. A more scalable approach utilizing the three sampling strategies is being tested in this study. The tests were conducted on only the four primary prediction predictors for prognosis because the performance of label grades in the initial phase was not as satisfactory. The literature suggests that the performance of classifiers is improved by making a more sensitive selection of representative of minority groups is made.

In this study, the most important labels include binary and multivariate categories of equal samples within each of the classes, since the size of the sample used in the study ranges from to 30000. The choice of ratios 50:50 for smaller samples are simple, resulting in results that will be weaker for larger samples. This proposed method for balanced has maintained the ideal threshold at 1:10, based on the data that is available from SEER-processed databases for all types of cancer as well as all the prominent cancer types. The specifics about the balance stratified method are given in Table 1. The algorithm was contrasted with traditional random sampling as well as stratified sampling to provide a better understanding.

Table 1 : Algorithm model

Algorithm	Balanced Stratified Sampling ratio
Input Output	Master data set D = {{c_1,x_1),(c_2,x_2) …..(c_n,x_n)}, class label C = {c_1,c_2…c_n}; size vector of all c_i in C Sample data set S<D; with size p
Step 1 :	Initialize the bin for each strata in SC={(sc_1, sc_2, ….sc_k} for c_i != 0 in D
Step 3 :	Find q=p/k ; the required size of each strata sc_i
Step 4 :	Add one random sample from each strata of D to each bin in SC

Step 5 :	Continue step 4 until p values are identified

IV. METHODOLOGY

The basic methodology is founded on Process Mining Project Methodology with the focus on mining and analysis

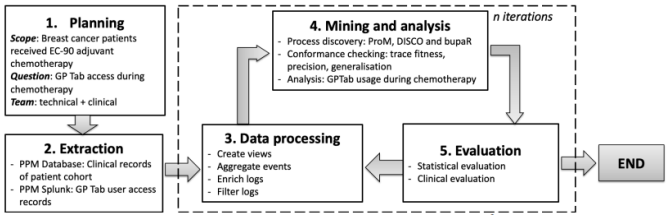

Figure 1. The general methodology

We completed the steps in the method in at least two times: once using solely the records from the clinic as an input, and another time with an amalgamation of the clinical records as well as the log of access to healthcare for users.

The planning stage determined the study's scope, the team and the research issues for the study.

The Extraction stage comprised the clinical information of patients taken from the PPM database, as well as the log of access for users from PPM Splunk.

The Data Processing stage involved creating views by aggregating events, adding value to logs, and then filtering logs. Views were constructed by looking at the chemotherapy cycles of patients with breast cancer. Instead of grouping events by date, we used the precise names for events, which include the Cycle 1 and Cycle 2 all the way to Cycle 6, which represents the number of cycles of chemotherapy. Log enrichment provided additional information to the log of events and, in this case, the duration of the process for each patient was determined as the days that passed from the beginning of the treatment until the end of the treatment recorded. We also added Emergency events as well as Neutropenia events, which were recommended by experts in the field as the two main events that could affect the progression of chemotherapy. We removed the Emergency events when they were recorded in the Admission table using the Emergency Type of Admission.

This stage employs the Neural PCA ensemble models to simulate the treatment method and create an exact model of the entire input scenario.

The Evaluation stage was carried out to determine, confirm and validate the findings of the previous steps. The evaluation was a review of all results from both the statistical and clinical perspective. The statistical analysis was conducted to confirm and validate the findings quantitatively. This was then confirmed by expert clinical experts as well as the representatives from the team responsible for development. The clinical evaluation was conducted to verify that the results reflected the actuality, and were backed by the existing knowledge of medical experts on patient care.

Data Extraction and Data Processing Stages:

We identified patients with cancer from the SEER Data which was used as adjuvant chemotherapy. The access to the GPtab was made available to doctors from 2014 until 2021. The number of patients was 738 in this grouping. Table 2 lists a summary of the eight events selected to determine the process, which comprises six cycles of chemotherapy as well as two adverse events.

The term "event log" also known as E is a sequence of events characterized by an event's case_id, a title, as well as a time stamp, (c, a, T). An event is a description of an activity that occurred in a particular instance with an exact time of that is t. Traces, or T is a sequence of events which occurred in a particular case, with the T is. In this research it is the case of the patient with an event sequence that occurs that occur between the time of the referral and diagnosis of endometrial cancer.

Table 2 : Events from the dataset logfile

Case_ID	Activity	Timestamp
P001	Referral	2020-01-06
P001	Investigation	2020-01-13
P001	Review	2020-01-17
P001	Diagnosis	2020-01-31
P002	Referral	2020-01-21
P002	Investigation	2020-01-22
P002	Review	2020-01-31
P002	Diagnosis	2020-02-10
P002	Surgery	2020-02-10

Sub-logs, S It is the sub-log of the event log E, according to the partitioning criteria. The partitioning must be performed in a manner that a trace is put into a sub-log, without duplicates among sub-logs. In this study this event log was split into sub-logs according to the diagnosis year for the patient. There are many options for partitioning that could be considered. The year of diagnosis was chosen as it was the primary stage in the selection criteria for patients in this study.

```
#Define Variables
c <- case_id
a <- activity
t <- timestamp
# diagnosis of patient
for each c,{
        Get YearDiag as year(t) where a = 'Diagnosis'
        }
# Create sub-logs based on year of diagnosis
for Year = min(YearDiag) to max(YearDiag){
        Create a sublog-Year
        Get all a of c having been diagnosed at YearDiag
        }
```

Analysis of these sub-logs relied on selected metrics at trace, model and activity levels in order to explain the multi-level nature of the relevant processes.

V. EXPERIMENT AND RESULTS

The data set comes from SEER from which we extracted time-stamped events occurring between referral and diagnosis of the patients we selected. This resulted into 339 distinct types of activity. Our study only focused upon the general categories of activities which are represented in the 9 tables. We separated discharges and admissions from the table that they were combined in and separated diagnostic surgery-related events from the table of surgery. The 11 activities that resulted were in agreement with the prior experience of our co-authors in the field of clinical research.

#	Activity Name	Occurrence (%)	Patients (%)
1	Referral	943 (12)	943 (100)
2	Diagnosis	943 (12)	943 (100)
3	Investigation	1455 (18)	891 (94)
4	Diagnostic Surgery	1025 (13)	797 (85)
5	Pathology	1196 (15)	540 (57)
6	Admission	661 (8)	285 (30)
7	Discharge	581 (7)	193 (20)
8	Consultation	346 (4)	128 (14)
9	MDT Review	338 (4)	199 (21)
10	Surgery	248 (3)	234 (25)
11	Outpatient	231 (3)	135 (14)

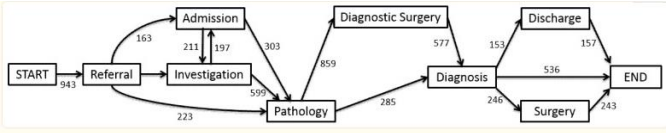

Figure 3 : Process flow

The model of the process route between referral and diagnosis. The pathway runs across the left from right to left with rectangles representing activities , and the arrows indicating how it flows between one activity and another. Arrows with numbers indicate the percentage of patients who have flow of activity into other activities.

RAndom K-labEL (RAkEL) subsets train m random subsets k labels by using combination models within an ensemble model using the threshold for producing predictions for the set of labels based on the collective votes of the m models during testing time. The complexity of RAkEL is therefore limited by k instead of in which k is greater than L. This technique has been one of the most widely recognized in the literature on multi-label and is used as a benchmark in various evaluations, including this study.

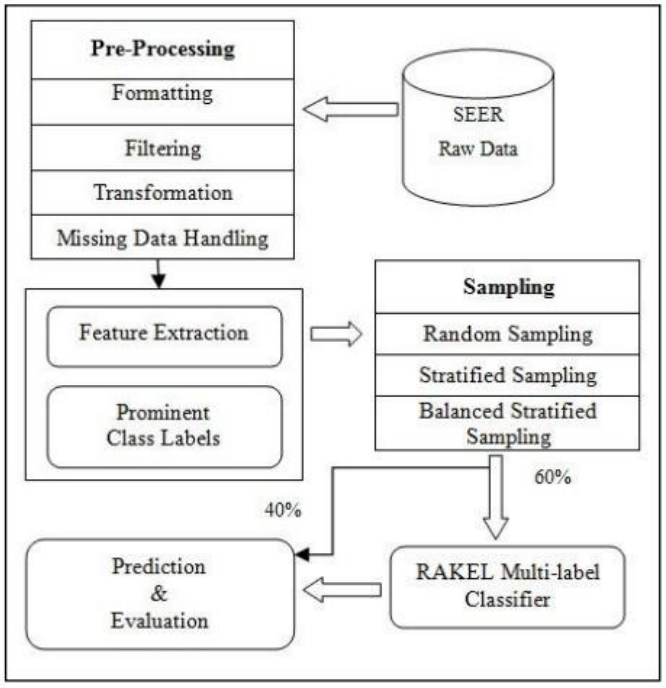

Figure 3: Model architecture

Metrics were compared and evaluated in a way that is relevant to those with cancer in the data collection. In the test, n fluctuates between 10000 and 30000 in five bins. MATLAB files were integrated with basic function files for the evaluations of the experiment [10]. The data sets extracted from the sampling chapter that leave the class label's age has been utilized for the research. RAKEL classifier can be described as an improved version of the ensemble algorithm that uses the ratio of class labels for the selection of samples in every iteration.

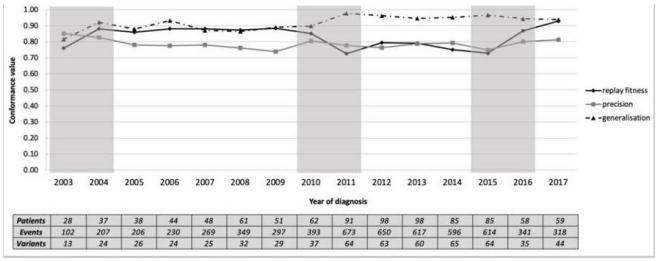

Figure 5 : Conformity to the annual process model. The areas in shaded form show times when changes could occur at the model level.

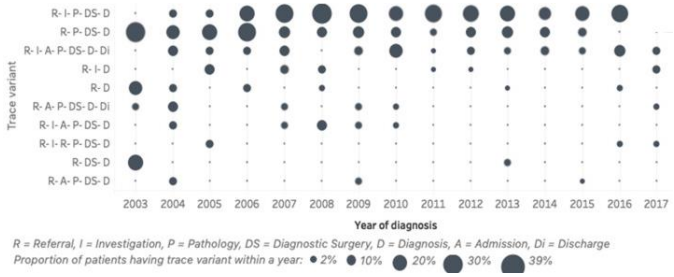

Figure 6: The summary below summarizes the trace variant comparision (2003-2017). Size indicates the proportion of trace variants compared to the number of patients diagnosed the year.

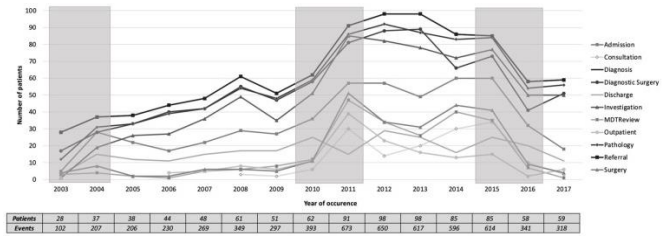

Figure : Total number of individuals who participated in every one of the major activities (2003-2017). The areas shaded show time frames during which changes could occur at an activity level.

The stage of extraction was when experts from the field evaluated and recommended specifics for the extraction process. A key suggestion at this stage was to select the kind treatment for cancer of the breast chosen in this study that is adjuvant therapy. In the stage of data processing clinical experts suggested that they focus on the effects on the introduction of chemotherapy cycles. The results of the mining and

analysis stage were discussed with experts in the field of clinical medicine.

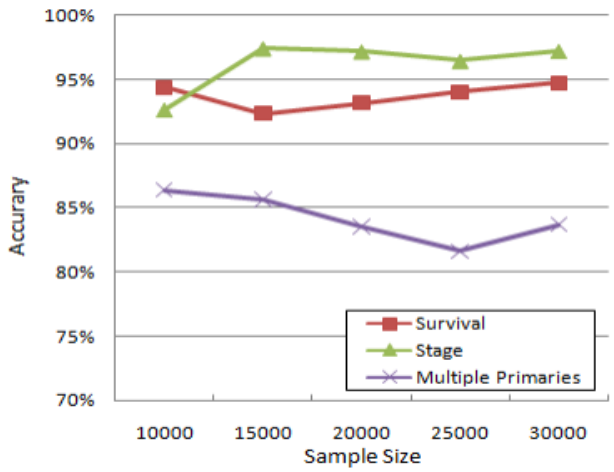

Figure : Accuracy of breast cancer data set

VI.CONCLUSION

The work proves to be efficient and accurate in terms of extracting the data from the health care HIS and HER . The works is proposed with noval extraction of the events from the data set and applying the neural PCA in order to reduce the dimensions of the events. Hence further ensemble model The RAKEL design is an intriguing and innovative approach because it is able to balance the characteristics of sampling as well as the increasing the class of selection. This research demonstrates a typical improvement in the efficiency of RAKEL over the SB in this breast cancer dataset, using three distinctive names for samples from five bins. The main drawbacks of RAKEL is mostly the amount of work involved in balancing samples as their size increases. The entire process is expected to become automated within the next few years the use of RAKEL to forecast the live data from the application.

VII. REFERENCES

[1] American Cancer Society, "The History of Cancer," 2011. www.cancer.net/patient/Advocacy and Policy/Treatment_Advances_Timeline.pdf (accessed Aug. 09, 2016).

[2] CRUK, "Your cancer type," Cancer Research UK, 2014. http://www.cancerresearchuk.org/about-cancer/type/ (accessed Aug. 09, 2016).

[3] National Cancer Institute, "Breast Cancer - Patient Version," 2018. https://www.cancer.gov/types/breast/patient/breast-treatment-pdq.

[4] Office of National Statistics, "Cancer registration statistics, England:2016," 2016.

[5] National Institute for Health and Care Excellence, "Advanced breast cancer overview - NICE Pathways," NICE Pathways, 2016. https://pathways.nice.org.uk/pathways /advanced-breast-cancer.

[6] National Chemotherapy Advisory Group, "Chemotherapy Services in England : Ensuring quality and safety," Dep. Heal., no. August, 2009.

[7] Royal Cornwall Hospitals NHS Trust, "Clinical Guideline for the Assessment and Management of Chemotherapy Induced Diarrhoea," pp. 1–11.

[8] W. M. P. van der Aalst, Process Mining: Data Science in Action, 2nd ed. Springer-Verlag Berlin Heidelberg, 2016.

[9] E. Rojas and J. Munoz-Gama, "Process mining in healthcare: A literature review," J. Biomed. Inform., vol. 61, pp. 224–236, 2016.

[10] R. Mans and W. M. P. van der Aalst, "Process mining in healthcare: Data challenges when answering frequently posed questions," Process Support Knowl. Represent. Heal. Care, pp. 140–153, 2013.

[11] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process mining in oncology: A literature review," in Proceedings of the 6th ICICM 2016, pp. 291–297, doi: 10.1109/INFOCOMAN.2016.7784260.

[12] A. P. Kurniati, E. Rojas, D. Hogg, and O. Johnson, "The assessment of data quality issues for process mining in healthcare using MIMIC-III , a publicly available e-health record database," no. 2, 2017.

[13] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, "Process Mining in Oncology using the MIMIC-III Dataset," IOP J. Phys. Conf. Ser. 971, vol. 971, no. 012008, pp. 1–10, 2018.

[14] A. Newsham, C. Johnston, and G. Hall, "Development of an advanced database for clinical trials integrated with an electronic patient record system," Comput. Biol. Med., vol. 41, no. 8, pp. 575–586, 2011.

[15] K. Baker et al., "Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy," Int. J. Med. Inform., vol. 103, pp. 32– 41, 2017, doi: 10.1016/j.ijmedinf.2017.03.011.

[16] Leeds Teaching Hospitals NHS Trust, "Leeds Teaching Hospital," 2016. http://www.leedsth.nhs.uk/ (accessed Jul. 26, 2016).

[17] W. Hazell, "Analysed: The biggest NHS providers of specialised services | News | Health Service Journal," 2015. https://www.hsj.co.uk/home/analysed-the-biggest-nhs-providersof-specialised-services/5091147.article (accessed Jul. 30, 2019).

[18] LTHT, "Leeds Care Records GP Tab," Leeds, 2019. [Online]. Available: https://www.leedscarerecord.org/lcr/widget/whats-in/gp-tab.pdf.

[19] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van Der Aalst, "PM2: A process mining project methodology," in International Conference on Advanced Information Systems Engineering, 2015, pp. 297–313, doi: 10.1007/978-3-319-19069-3_19.

[20] O. A. Johnson, T. B. A. Dhafari, A. Kurniati, and E. Rojas, "The ClearPath Method for Care Pathway Process Mining and Simulation," in Lecture Notes in Business Information Processing, 2018, pp. 1–12.

[21] A. Adriansyah, "Replay a Log on Petri Net for Conformance Analysis plug-in.pdf | Algorithms | Logarithm." Accessed: Oct. 31, 2017. [Online]. Available: https://www.scribd.com/doc/225913652/Replay-a-Log-on-PN-for-Conformance-Analysis-plug-in-pdf.

[22] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van Der Aalst, "On the role of fitness, precision, generalization and simplicity in process discovery," in OTM Confederated International Conferences, 2012, pp. 305–322, doi: 10.1007/978-3-642-33606-5_19.

[23] W. M. P. van der Aalst, A. Adriansyah, and B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," Wiley Interdiscip. Rev.Data Min.Knowl. Discov., vol. 2, no. 2, pp. 182–192, 2012, doi: 10.1002/widm.1045.

[24] W. M. P. van der Aalst, B. F. Van Dongen, C. Gunther, A. Rozinat, H. M. W. Verbeek, and A. J. M. M. Weijters, "ProM: The process mining toolkit," CEUR Workshop Proc., vol. 489, 2009.

[25] C. W. Günther and A. Rozinat, "Disco: Discover Your Processes.," in BPM 2012 Demonstration Track, 2012, vol. 940, pp. 40–44.

[26] G. Janssenswillen, "bupaR: Business Process Analysis in R," R package version 0.4.2, 2019. https://cran.r-project.org/package=bupaR.

Cite this article as :

Smt. S. Yamuna Rani, Dr. Sumagna Patnaik, " Rakel Model For Multi Class Label Classification Using Ensemble Neural PCA On Healthcare Event Log", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 334-342, November-December 2022. Available at doi :
https://doi.org/10.32628/CSEIT228646
Journal URL : https://ijsrcseit.com/CSEIT228646 |