

Systematic mapping in improving the extraction of Cancer Pathology information using RPA orchestration

Sreekrishna M,* Prem Jacob T

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India.

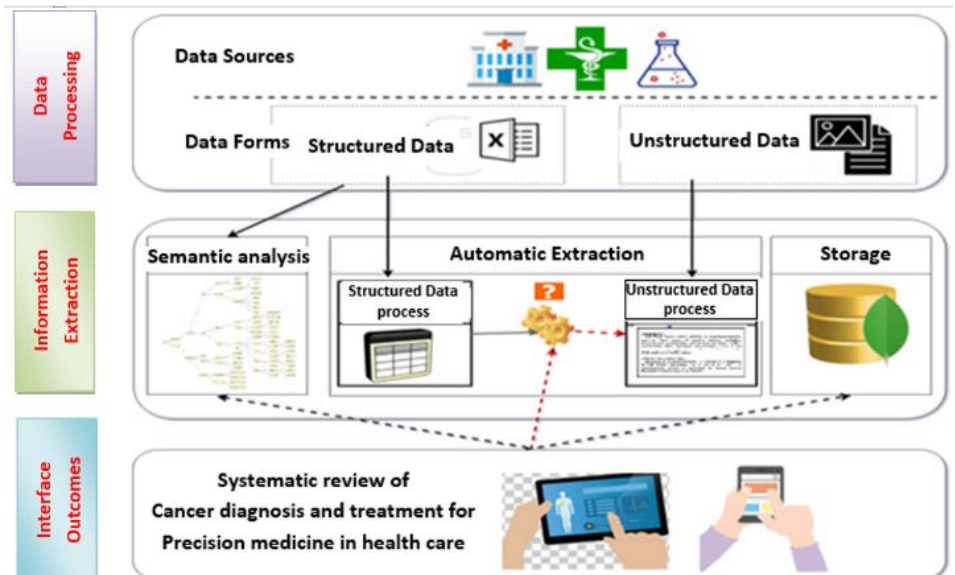
Received on: 12-Jan-2023, Accepted and Published on: 2-Apr-2023

ABSTRACT

Integrating cancer medical data from various sources has significant impact in analysis and treatment of cancer. An efficient methodology is introduced to extract quantitative information from the unstructured pathological data. Computational intelligence in Robotic Process Automation (RPA) is done to process the data and extract the information for evaluating patients from clinical pathology report. The observational study has been undertaken to examine the factors influencing the analysis of cancer in order to improve treatment process. The data

has various stages of information like patient information, Grade of cancer, medical history and treatment undergone. Consideration is given to automate 24 unique features from 767 individuals. The pathologist considers particular terms for data extraction to specify the Natural Language Processing (NLP) semantic rules. On the chosen specimens with 24 data items, the suggested RPA NLP technique obtained accuracy scores of 98.67%, recall as 98.55% and F-measure scores of 98.45%. The feasibility and precision of autonomously extracting pre-defined data from clinical narratives for cancer research has been established in this work. Thus, Enhanced RPA implemented automates repetitive tasks by analyzing daily interactions for cancer patients who receive personalized care.

Keywords: Unstructured data, Intelligent automation, Natural Language Processing, Feature analysis, Prediction, Supervised Learning



INTRODUCTION

Cancer is caused by the growth of abnormal cells which are classified as benign, malignant, or semi-malignant and is a leading

cause of deaths worldwide.¹ Prompt detection is required for better treatment and precise medicine. Earlier, prediction is primarily based on clinical characteristics such as age, tumor shape, features and surgical characteristics.² Physicians have recognized that the same disease can manifest itself differently in different patients and that there is no one-size-fits-all treatment. Precision cancer medicine improves cancer analysis and diagnosis of cancer to process with the most effective treatments.³ Effective treatment can be provided by the healthcare professionals by mapping with the existing relevant tumor mutated patients. Clinicians are increasingly responsible for gaining information from patients in

Corresponding Author: Sreekrishna M, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, India: Email: krishsree1212@gmail.com

Cite as: *J. Integr. Sci. Technol.*, 2023, 11(4), 561.
URN:NBN:sciencein.jist.2023.v11.561

©Authors CC4-NC-ND, ScienceIN ISSN: 2321-4635
http://pubs.thesciencein.org/jist

the digital day for managing cancer health data.⁴ Patients have the right to access, analyze, and share their substantial changes in their health. After sharing self-tracking data, the majority of patients are unsatisfied with their health care providers. By comparing cancer patient health data into current health data systems, it is still possible to improve prognosis of cancer health care.⁵ Various types of cancer health information have been recognized in the literature. Medication information, population-based registry, special cancer registry, social interaction data, genetics, psychological data, symptom data and reports are among these categories. Despite this, only a few studies have looked into how individuals' health data is stored. According to studies, health data management is a major concern.^{6,7} The huge amount of health data must be managed by many different sorts of organization. The state registry of cancer holds and maintains the characteristic features of cancer like origin, species that are collected by the National Program of Cancer Registry,⁷ which supports in making educated decisions by the healthcare professionals. These registries serve variety of purposes, that includes managing the information of cancer history, identifying population-based cancer patterns based on its classification, awareness in providing cancer care programs for controlling the cancer, progressing in supplying information for a cancer care registry and doing medical, statistical, and health assessment. Manage the systematic gathering, sharing, and processing of cancer data seems to play a major hub for American cancer research. Traditionally, a qualified cancer recorder records collected information into cancer registries, who used an electronic or manual technique to document the information into electronic or paper forms. The process of extracting data from cancer registry starts with initial examination and diagnosis of a patient. After that, the information is entered into a healthcare register database.⁸

The medical content in a variety of formats, that includes emails, texts, reports and documents. The majority of this information is unstructured, with only defined formats, and it can be difficult to extract information from unstructured language. Text-based unstructured data is managed and analyzed using natural language processing.⁹ The backbone of healthcare for data analysis is the use of natural language processing (NLP) methods that converts unstructured data from various sources into standard data suitable for computational analysis. NLP, a set of methodologies used to analyze text from unstructured data, has made great progress. Extraction of data basically classified into two main categories. If the relation to be extracted is predefined then it is known as traditional information extraction.¹⁰ The relation when not defined initially and the system is unrestricted in how it extracts relationships from the text data. Extracting information manually is not efficient when the volume of data increases. Automation in data extraction can be done. We construct a set of rules for the syntax and other grammatical aspects of a natural language. In Supervised Learning algorithms that accurately identify data or forecast outcomes are trained, which makes use of labelled datasets. Semi-supervised learning creates unique patterns that can be utilized to extract more relations from the text. In order to train the model, both labelled and unlabeled data are used.

The primary problem in the healthcare sector is managing and processing patient information in the healthcare that takes

significant amount of time to accomplish different systems with administrative and rule-based tasks. As a result, patient services are slower, resulting in poorer productivity and increased wait times. In the research area of Machine learning research, natural language processing can be implemented that can extract the data from all types of text documents. Text sequences are analyzed at the initial stages, which are known as parsing of Natural Language.¹¹ A semantic interpretation of meaning based on the texts is also introduced as a result of information extraction. The goal of this research is to create and evaluate strategies for extracting relevant data from clinical texts with a focus on treatment for cancer. The goals are to develop tasks that are time-consuming in the hospital and assess the potential of RPA in order to design an application, in which cancer registry dataset can be updated automatically

RESEARCH BACKGROUND

Pathology reports are the main source of data for diagnosis of cancer. The periodic clinical information in pathology reports, specifically focus on free text, making it challenging to access data.¹² Several NLP methods were recommended for extracting the text automatically from malignancy report of the cancer. An accurate intraoperative diagnosis for cancer patients is highly required for personalized treatment. The goal of this work is to create an automated keyword extraction algorithm to summarize the existing narrative pathology report into a more focused pathology report. The observational study has been carried out to examine factors influencing the pathology report analysis in order to improve the treatment. For automatic extraction of features it is necessary to find, analyze, and obtain all the information from cancer pathology reports as shown in Figure 1. The documented medical information on cancer can be used as a knowledge resource to improve healthcare.

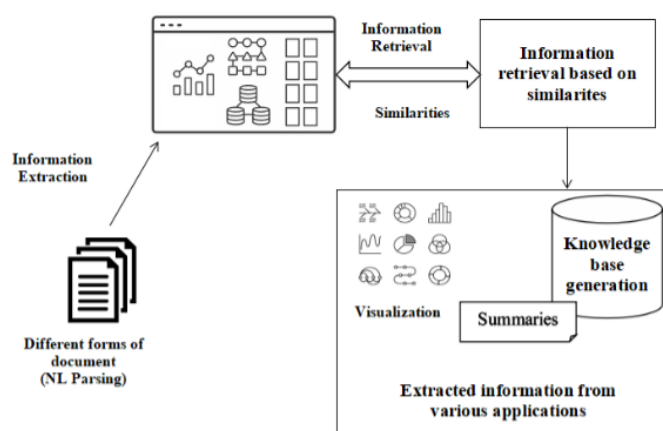


Figure 1: Implementing RPA NLA for text extraction

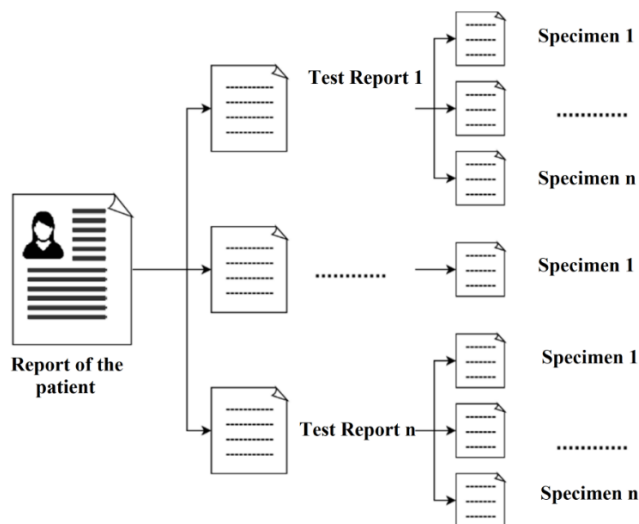
Clinical text information is considered to be a form of unstructured data that is challenging for algorithms to automatically analyze.¹³ NLP and textual categorization are widely used in the fields of biomedicine and health informatics. The research is to create and evaluate strategies for extracting relevant data from clinical texts with a focus on treatment for cancer. The implementation of structured data helps to facilitate information

sharing among organizations, recent initiatives have looked at better ways of managing pathology reports. Approaches have been made for standardizing cancer reports and structured data is being encouraged to create functional requirements that define the system for pathology reports. The investigation and processing of massive text data require quick, accurate tokenization and information retrieval from free-form texts in natural language

LITERATURE REVIEW

In 2020, 48,639 cases of cancer was registered and diagnosed in Malaysia. Based on histopathology, each confirmed cancer diagnosis is supported by a pathology report. Specimens are the cells and tissues that which are documented in a pathology report.^{14,15} These specimens are prepared by a pathologist depending on microscopic information and are used to diagnose disease. Clinicians can pick the best course of therapy for a patient by reading the report's description of the tissue to determine whether it is malignant or not. The macroscopy, microscopy, and gross description are the preferred standard conventional narrative pathology information. These reports serve as a valuable resource for information on certain tumor traits. Unfortunately, traditional reporting of tumors in free-text format is sometimes accompanied by complicated explanations and inconsistent terminology. As a result, practitioners who are familiar with the contents of pathology reports must manually extract the necessary details, from these reports.¹⁶ Manual conversion of periodic reports to standardize the information can be expensive and consumes more time, the reports quality can be improved by synoptic reporting. Ongoing research is automatic transformation of the documented test result which are automated as synoptic format, where it employs NLP that can detect relevant data items from documented pathology reports and obviates the need for human intervention. NLP has recently been popular in the medical fields, particularly for retrieving detailed info in electronic patient medical details with accurate accuracy ranging between 86% and 98%. The National Population-based Cancer Program concentrates on cancer relevant data from the general population, while the National Hospital-based Cancer Program concentrates on cancer-based classification from inpatient cancer registries.^{17,18} Traditionally, a certified cancer registrar entered data into cancer registries by using a manual or electronic method to record the details into electronically or in paper forms. The initial patient evaluation and diagnosis with the beginning of the data extraction procedure from a cancer registry. After that, the information is entered into a healthcare register database. A cancer registry at a hospital collects data from on all cancer patients who obtain care from a health-care provider. To keep the patient's status up to date, a phone call was done annually. A cancer registry is a centralized repository for cancer data gathered by cancer registrars. Every new patient admitted to their facility for cancer-related treatment is reported by each healthcare provider.¹⁹ The development of NLP tools to amplify and automate extracting of information from tumor histology reports is the issue plaguing the cancer SEER surveillance program. The annotation done manually is currently which is not highly adaptable to gather all data items required to observe the continuous tales of affected patients due to high volume of tumor pathology reports that are produced annually

utilizing cutting-edge NLP technologies.²⁰ In order to find significant data features from pathology reports semantic information approaches have been adapted for medical tasks, such as categorization of information based on its features and extracting relevant information, as systematic learning approaches for variety of NLP tasks for parsing, sentiment analysis, and question answering. Previous research on the extraction of data from clinical information addresses a wide range of tasks and research problems.²¹



Each specimen has its own interpretation section.

Figure 2: Pathology report representation

For surgical pathology reports, caTIES supports coding. The hierarchical model identifies and utilizes a regular expression to analyze histopathology information to depict cancer diseases samples and related data. Lupus uses Semantic Web tools to express extracted data. For the purpose of annotating surgical pathology reports, NegEx is used to identify negation. These training data that are annotated use rules that are tailored to particular subjects and domain methods that have been learnt individually. The method cannot be generalized, new rules must be created for every domain.²² In order to train a supervised neural NLP model, a sizable training dataset with comparable properties as the prospective sampling dataset. On the other hand, acquiring a substantial labelled clinical text database requires a lot of time and effort. Consequently, a variety of methods have been suggested to address hoe to overcome practical constraint like learning while multitasking and adaptive learning. In narrative free text, pathology reports incorporate useful research information. Despite a significant movement towards organized reporting, most pathology reports in legacy systems are still unstructured as shown in Figure 2. Additionally, standardization initiatives only capture the main data pieces, leaving a sizable amount of important data in textual information that is challenging to analyze and search.²³

Recent advances in NLP have taken steps towards automating textual categorization and retrieval of information from reports. The diagnosis process in pathology necessitates the compilation of

data from various clinical and pathological sources which is a challenging intellectual process.²⁴ To develop machine learning pathology methods for high performance diagnostic analysis, datasets must be scalable annotated with semantic labels from corresponding synopses. It can be challenging to acquire semantic labeled data using deep learning and machine learning models. By investigating the required labeled data is necessary to train a model successfully, active learning approaches may be able to solve this issue. In this study, Convolutional Neural Network is adopted as the text classification model to compare the performance of active learning algorithms on categorizing subsite and histology from cancer pathology reports. Given the availability of several clinical tests, designing a specific framework to address every scenario that can be seen in the pathological analysis is difficult.²⁵ It is challenging for clinicians to spend time utilizing various standards to record the information. The technology will ideally be capable of automatically retrieving the information that the pathologists require by using NLP as input. In order to extract information from various histopathology, research on turning unstructured histopathologic textual data into a format data is becoming a necessary requirement. In order to achieve customized oncology, technological advancements have made it possible to detect an abundance of molecular biomarkers. Large clinical series with high-quality, well-annotated data are crucial for personalized cancer research. Utilizing extensive, real-world practice data recorded in EHRs to support translational and clinical research is one emerging area of study. In EHRs, like as pathology reports, a lot of this data, such as tumour features, may only be accessible in descriptive or semi-structured language. Despite the pathology community's movement toward synoptic reporting, this formatting has not yet been fully adopted and panoramic reports are often only produced after definitive resections. In the medical field, namely the cancer sub-field, NLP techniques that can extract features and arrange information from textual documents have been thoroughly studied. Information extraction recognition, a crucial NLP approach that locates and assigns named things to predefined categories, has received a lot of attention in studies. Numerous NLP systems have been created to handle pathology reports which are rich with details regarding the features of tumor specimens. Information is previously extracted using pathology reports to promote cancer research.²⁶ In addition to information retrieval and idea coding capabilities, described the Tumor Tissue Information Extraction Method, which concentrates on de-identification of tumor tissue for research purposes. Anatomic location, histology, and grade are among the cancer features, a rule-based system, intended to extract from pathology reports.²⁸⁻³⁰ Recently, a rule-based approach was developed to extract information on cancer stage, morphology, and topology from clinical data. Although cancer feature extraction using clinical text has been successfully documented use cases, NLP has not yet extensively applied for regular cancer research, most likely because of implementation difficulties.^{31,32}

Due to the complexity and diversity of clinical information, developing specific medical NLP systems or extending current systems for particular applications requires a great amount of work. The correctness and comprehensiveness of the data continue to be

the fundamental barriers to the processing and analysis of clinical texts. Clinical NLP solution that offers both user-friendly interfaces and high-performance NLP components for creating unique NLP pipelines for every client.^{33,34} Efforts are taken to create a collection of feature extraction modules for critical tumor information in pathology reports in this paper. Users can quickly construct customized NLP systems to retrieve cancer information from local pathology reports using such built-in modules and data interfaces. It has been acknowledged that automated encoding of information from free-text histopathology as a helpful tool for identifying rare cancer findings for tumor registration. Information has been reliably and accurately extracted both free-text histopathology and from patient narratives for malignancies using a variety of machine learning techniques. Natural language processing has at least been used in one study to extract tumors from test results, although the system did not extract diagnostic or location information. Currently, a number of ML based methods have been created and put to the test in histopathology to help with pathological diagnostics using the fundamental morphological pattern.³⁵ A kind of machine learning called deep learning uses artificial neural network (ANNs), in which the statistical models are built from training data input. Similar to a biological deep neural network, ANNs can independently decide whether their analysis is accurate. Clinically NLP solutions are using procedures that depend on a list of procedures created by experts that specify how a machine should categorize a document. Creating rule-based solutions takes a lot of effort considering that there are numerous methods to communicate the relevant information. However the existing methods are not efficient in extracting quantitative cancer data.

RESULTS AND DISCUSSION

The major goal is to provide time-consuming actions and assess their potential in order to design an application in which a software bot process to complete the operation efficiently. The RPA application will be implemented using the UiPath tool. Conceptual machine learning is assessed to expand the RPA application usage.^{26,27} Since RPA solutions cannot be created for every area of work in the healthcare industry, tasks are determined that are appropriate for RPA implementation. The time-consuming and rule-based manual tasks performed by healthcare personnel are investigated. People come to see RPA as an invaluable digital helper when it is at its most effective. Following the analysis of the data, a few use cases are discovered that are suitable for RPA deployment, appointment Scheduling based on Patient's Health Record. Introducing additional patient knowledge to the current document and searching the document for prognosis-related keywords.

The orchestrator is used to schedule and keep track of the robots as they operate the various machines installed automation processes., while the robot is used to execute the processes. The sorts of project templates accessible in the UiPath studio is shown in Figure 3 to construct a blank automation project based on the requirements of the user. Developing code snippets and making them accessible as a library are two ways to achieve this. The library can also be used in other automation program as a

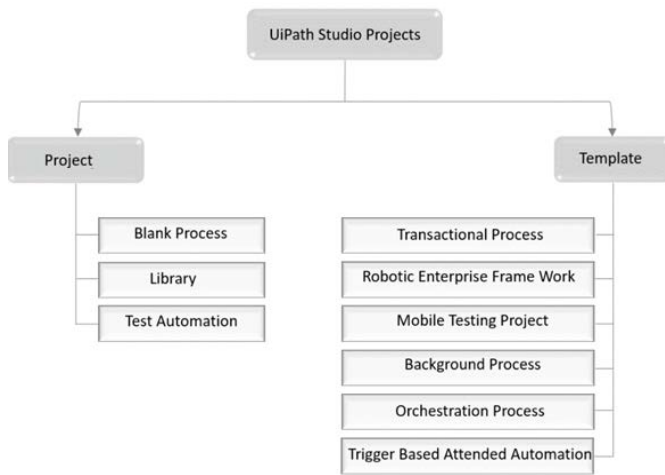


Figure 3: Importing of data into UiPath

dependency. Orchestrator's Test Automation feature allows you to deploy and manage your tests. Excel, mail, credentials, system, and UI automation are the default activities in this project template. With an orchestration service, lengthy workflows and human input are all automated using this template. If the additional purposes does not require user interface activities, this framework is used to construct a predecessor that runs concurrently with some other activity on the same robot. The Activities pane contains a list of tasks to process. To interconnect activities in the workflow, the properties of each activity is set up in the properties pane. The Output pane displays process execution data along with project dependencies. UiPath studio is made up of a number of activities that are combined into packages. A few default programs are displayed as dependencies whenever a new application is started in the studio. These packages comprise spreadsheet, mail, system and web application operations. The RPA application is also developed using database and PDF operations.²⁷ To execute processes and establish a link with the orchestrator, the robot must be installed in the studio. The robot executes the automation projects in UiPath studio, as shown in Figure 4.

An EBHR contains a patient's information and details such as identity, age, and location, as well as health - related information, like lab testing results, X-ray, scans, and diagnostic test reports. The patient's medication information is stored in the MetaVision clinical information system. The conceptual system is used to record the prescription and dosage for the patients. The dataset "Cancer Data" from Gitlab catalogue is utilized in this paper to train the bots. There are 767 instances each with 24 properties, one dependent variable and twenty-four independent variables. The information is most effective since integer values are typically present for comparative analysis. The dataset, support in manipulating the level of cancer by comparing with the attributes as follows. Using a dataset of anatomic histopathology from 767 patients, the system elements are tested to extract data such as Patient ID, Passive Smoker, Chronic Lung Diseases, Age, Dust Allergy, Swallowing Difficulty, Balanced Diet Chest pain, shortness of breath, fingernail clubbing, alcohol use, gender, workplace dangers, obesity, Blood Cough, Wheezing, Frequently Catching a Cold, Air Pollution Index, Genetic Risks, Smoking, Tiredness and Dry Cough. The significance Levels of High, Medium and Low correspond to benign tumors, malignant tumors and healthy individuals free of lung cancer, respectively and the data and its impact are visualized in Figure 6. The annotation recommendations are analyzed and also evaluates the annotated data to resolve any issues. The seed phrases made up of lexical items for diagnosis, genes, and procedures are used that we collected from the Sentient Disease Ontology, the Cell Cycle Ontology, and the NCI Thesaurus Ontology, respectively, to enable extraction.

Reading Patient Medical Record from the Document

The pathology form for the individual, which includes personal details, health information as well as other pertinent data, is then captured with an Optical character recognition device and exported as a PDF file by the healthcare personnel. Both patient information and clinical records are extracted by the bot during this automation procedure from the document of the medical form and saved in an excel sheet. Utilizing pre-built activities, UiPath Studio provides an interactive designer for creating both simple and complex automation processes. Manage Packages under the ribbon tab can be used to install additional activity packages in the studio. This design includes a sequence of phases that enabled the automated reading of cancer detailed treatment forms as shown in Figure 5. The following are the step in creating the process flow for processing medical documents is the development phase in UiPath studio by utilising tasks. The pdf file on the file directory is opened with the support of 'For Each' activity. Personal information entered by the patient is read via activities such as 'Anchor Base' and the searching is processed with 'Find Element,' and 'Get Text.'

The "Find Matches of the Image" activity reads the option name if the patient has marked it in order to collect the health history as shown in Figure 5. Information about the patient taken from the detailed treatment form is shown in Figure 6.

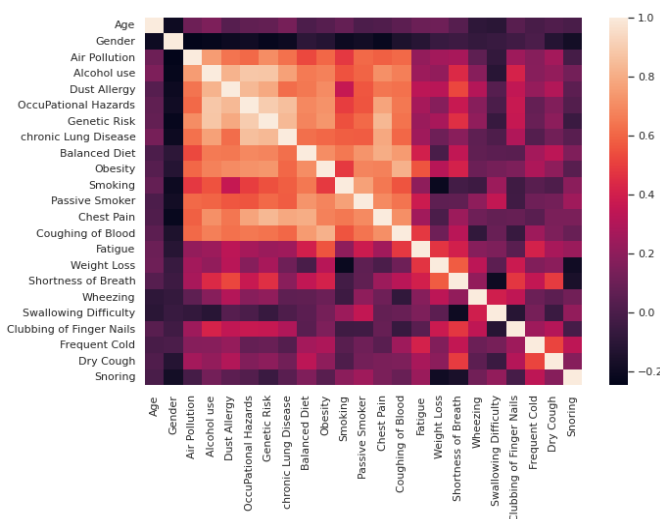


Figure 4. Heatmap of data visualization

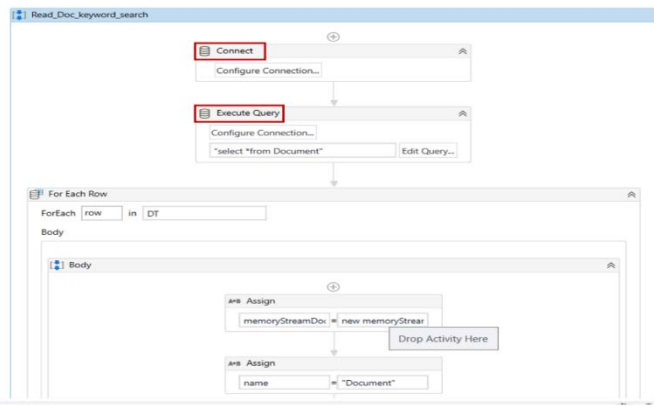


Figure 9: UiPath Studio method for automating keyword search

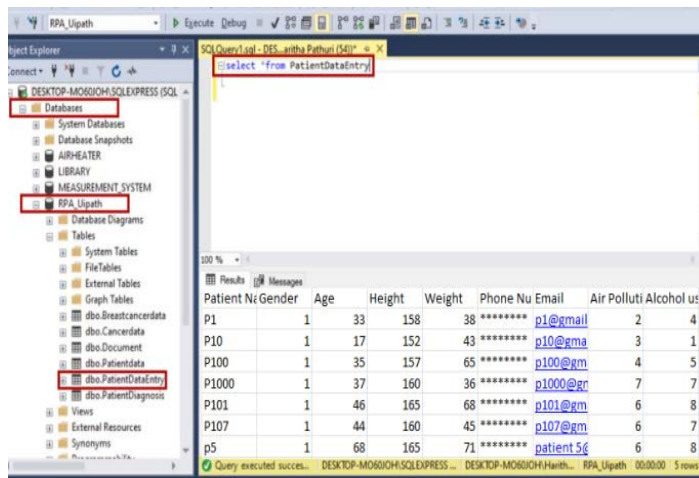


Figure 10: Patient Diagnostic information

database on the SQL server is where these two tables such as Report and Patient Diagnosis, are created.

Connecting the data to orchestrator:

Several computers in the network can be used to run the automation procedures that have been deployed to the orchestrator. The orchestrator must be made aware of the automated processes for this to happen. To accomplish this, Out of the device manager to the scheduler, the UiPath bot. Log in to the web application orchestrator as in Figure 11. Utilize the machines option in the Management menu to add the local machine of the orchestrator.

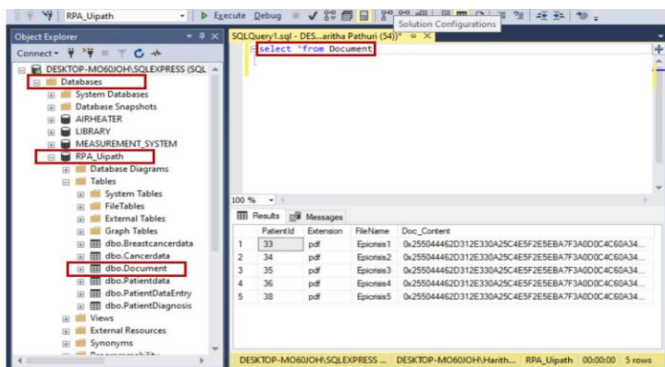


Figure 11: Connecting Patient Pathology information to orchestrator

The machine identifier is only accessible after the machine name has been established. From the system tray, access the UiPath robot's orchestrator settings as shown in Figure 12. Enter the machine key and the orchestrator URL. By selecting the link button, The orchestrator can be connected to the automation bot. The orchestrator will be linked to the Automation studio when the UiPath robot tray is connected. Now, the UiPath studio may publish the automation processes directly to the orchestrator. Automation processes can be installed on the Procedure tab of the orchestrator by choosing the add option and then entering the authored sequence of events into the Package Name box in the orchestrator Interface. To conduct the implemented process in the orchestrator, an environment must be constructed in the user interface (UI) of the orchestrator for grouping the robots and deployed processes. The Application server UI for displaying the names of the available robots and their environments is shown in Figure 13.

In BOTS tab, add the robot. Here, the computer's current machine name, robot name, username, and password, as well as the robot's kind, must be entered. The machine key can be copied from the MACHINES tab.

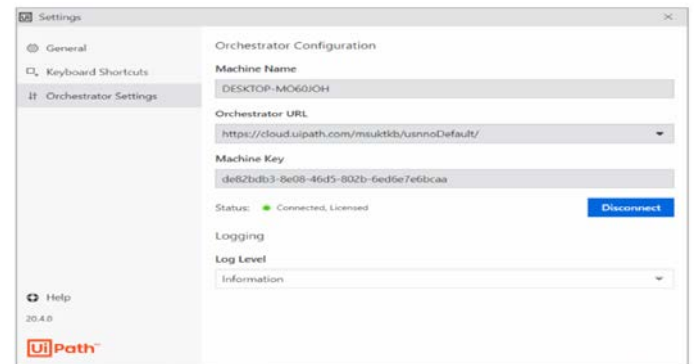


Figure 12: Settings for the orchestrator in the UiPath bot system tray

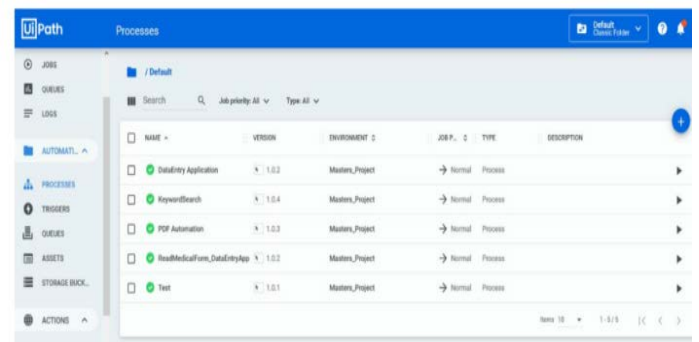


Figure 13: Deployment of Automation in Orchestration

The orchestrator's tab is used to select the process and assign the bot to the task to begin the execution.

Figure 14 depicts the beginning of the automated keyword search procedure using the designated robot. The machine that the robot has designated will be used to carry out the process.

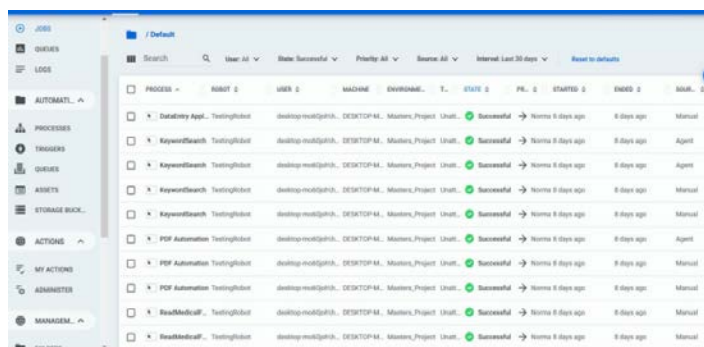


Figure 14: Automation process executed

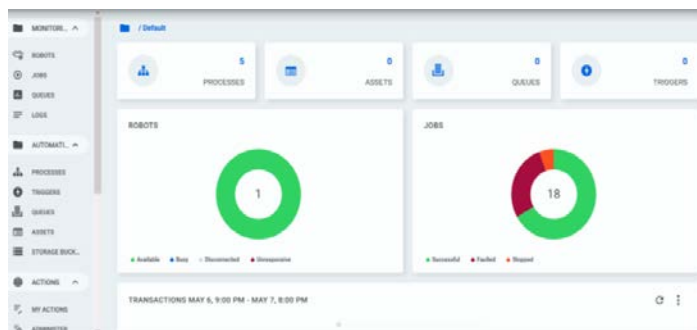


Figure 15: Processes deployed

Using the TRIGGERS in the orchestrator, bots are programmed for carrying out the sequence on a specific date or at a specific hour. Additionally, the orchestrator records every robot execution's specific. These specifics will be displayed in the Monitoring menu's LOGS tab. Figure 15 displays the orchestrator's active automation operations. The dashboard of the orchestrator shows important information like the number of processes running and the number of bots available.

Performance Metrics:

The efficacy of three upgraded supervised keyword extraction strategies are tested for various features from pathology report results, which were precision, recall and F-measure as shown in Table 1.

Table 1: Performance matrix for the supervised keyword extraction

Attributes	Precision	Recall	F-measure
Patient information	98.9	98.8	98.6
Symptoms	98.5	98.7	98.3
Diagnosis	98.6	98.1	98.2
Genetic marker	98.7	98.6	98.7

CONCLUSION

The therapeutic data in cancer is growing and software bots developed can make data management more accurate by eliminating errors. The used cases that are identified for the patients includes scanning the medical record to schedule an consultations, entering patient medications and treatments into the documentation, looking for specific diagnostic search terms in the document, providing medical terms, calculating the patient's prescription dosage in Meta Vision, and more. examining health records, inserting patient data into patient identification system and looking for keywords relating to diagnoses in an epicrisis document. The online orchestrator received the generated RPA apps and a robot is assigned to carry out the computerized use case processes. Bot retrieved patient information and information about the patient from the medical form's manuscript in reading patient medical form automated process and kept the data in an excel spreadsheet. The patient's information and medical history are read by a robot during the automation process from an In the aforementioned use case, a Spreadsheets document is created and inserted into patient registration application. The robot initially made contact and the diagnostic document is obtained using the Db database and then searched for using the diagnostic search terms in it. Later, these keywords and the patient ID are recorded in a separate table in the distant database. The suggested automated information retrieval approach for pathology reports based was authenticated over performance evaluation utilizing automated health data and useful extraction of keyword from annotated reports. With existing competing algorithms and appropriate results that comprise of accurate keyword extraction from misleading information, the current system demonstrated a significant performance. The biomedical researchers or medical organizations would utilize this work to address similar issues. Radiotherapy treatment planning and information tracking can also be done efficiently. The best outcomes can be used to improve treatment processing.

ACKNOWLEDGMENTS

The authors would like to express their sincere thanks and appreciation to the Sathyabama Institute of Science and Technology (SIST), India for their continuous encouragement and support.

REFERENCES AND NOTES

1. B.S. Chhikara, K. Parang. Global Cancer Statistics 2022: the trends projection analysis. *Chem. Biol. Lett.* **2023**, 10 (1), 451.
2. S.L. Goldenberg, G. Nir, S.E. Salcudean. A new era: artificial intelligence and machine learning in prostate cancer. *Nat. Rev. Urol.* **2019**,16, 391–403.
3. J.R. Strigley T. McGowan, A. Maclean, et al. Standardized synoptic cancer pathology reporting: A population-based approach. *J Surg Oncol.* **2009**, 99,517–24.
4. A.J. Gill, A.L. Johns, R. Eckstein, et al. Synoptic reporting improves histopathological assessment of pancreatic resection specimens. *Pathology.* **2009**, 41,161–7.
5. R.R. Karwa, S.R. Gupta. Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. *J. Integr. Sci. Technol.* **2022**, 10 (2), 110–119.
6. V. Jouhet, G. Defossez, A. Burgun, P. Le Beux, P. Levillain, P. Ingrand, et al., Automated classification of free-text pathology reports for registration

- of incident cases of cancer, *Methods of information in medicine*, **2012**, 51(3), 242.
7. K.C. Lee, B. Orten, A. Dasdan, W. Li. Estimating conversion rate in display advertising from past performance data, in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '12. New York, NY, USA: ACM. **2012** 768–776.
 8. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, Natural language processing (almost) from scratch, *J. Machine Learning Res.*, **2011**,12, 2493-2537.
 9. P. Khosravi, M. Lysandrou, M. Eljalby, et al., A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion , *J. Magnetic Resonance Imaging*. **2021**, 03.
 10. F. Schroeck, O. Patterson, Development of a natural language processing engine to generate bladder cancer pathology data for health services research, *Urology*. **2017**, 84-91.
 11. A. Wieneke, E. Bowles, D. Cronkite, et al. Validation of natural language processing to extract breast cancer pathology procedures and results, *J. Pathology Informatics*. **2015**, 07, 38.
 12. G. Napolitano, A. Marshall, P. Hamilton, A. Gavin, Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction, *Artificial Intelligence in Medicine*. **2016**, 70.
 13. M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, et al., Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J. Amer. Med. Inform. Assoc.* **2020**, 27(1), 89-98.
 14. Bray F and DM. Parkin. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability validity and timeliness, *Eur J Cancer*. **2009**,45, 747-55.
 15. Fan R.E, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification, *J. Mach. Learn. Res.*, **2008**, 1871–1874.
 16. J. X. Qiu, H.-J. Yoon, P. A. Fearn and G. D. Tourassi, Deep learning for automated extraction of primary sites from cancer pathology reports, *IEEE J. Biomed. Health Inform.*. **2017**, 22(1),244-251.
 17. T. Joachims. Text categorization with support vector machines: Learning with many relevant features, in Machine Learning: ECML-98, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, **1998**, 137–142.
 18. M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, et al., Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks, *J. Am. Med. Inform. Assoc.*. **2020**, 27(1), 89-98.
 19. A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, et al., Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *J. Biomed. Informat.* **2009**, 42(5),937-949.
 20. A.E. Johnson, T.J. Pollard, L. Shen, L.H. Lehman et al. MIMIC-III, a freely accessible critical care database. *Scientific data*. **2016**, 3.
 21. A. Esteva, B. Kuprel, R.A. Novoa et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. **2017**, 542(7639), 115–118.
 22. V. Gulshan, L. Peng, M. Coram, M.C. Stumpe et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. **2016**, 316(22), 2402–2410.
 23. J.A. Strauss, C.R. Chao, M.L. Kwan, S.A. Ahmed, J.E. Schottinger, V.P. Quinn. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J. Am. Med. Inform. Assoc.*, **2013**, 20(2), 349–355.
 24. S. Borde, V. Ratnaparkhe. Optimization in channel selection for EEG signal analysis of Sleep Disorder subjects. *J. Integr. Sci. Technol.* **2023**, 11 (3), 527.
 25. F. Sebastiani. Machine learning in automated text categorization, *ACM Comput. Surv.* **2002**, 34(1), 1–47.
 26. J. Ribeiro, R. Lima, T. Eckhardt, S. Paiva. Robotic Process Automation and Artificial Intelligence in Industry 4.0 – A Literature review. *Procedia Comput. Sci.* **2021**, 181, 51–58.
 27. T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In: *Adv. Neural Inform. Processing Systems*. **2013**, 3111–3119.
 28. Aguirre, A. Rodriguez. Automation of a business process using robotic process automation (rpa): A case study, in Workshop on Engineering Applications. Springer. **2017**, 65–71.
 29. S. Gehrman, F. Dernoncourt, Y. Li et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, *PLOS ONE* **2018**,13(2).
 30. S.N. Kakarwal, A.S. Gavali. Context aware human activity prediction in videos using Hand-Centric features and Dynamic programming based prediction algorithm. *J. Integr. Sci. Technol.* **2022**, 10 (1), 11–17.
 31. G.K. Savova, J.E. Olson, S.P. Murphy, V.L. Cafourek et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *J. Am. Med. Inform. Assoc.* **2012**,19(e1), e83–e89.
 32. A. Vaswani, N. Shazeer, N. Parmar et al. Attention is all you need. *Adv. Neural Information Processing Systems*. **2017**, 5998–6008.
 33. R.I. Doğan, R. Leaman, Z. Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J Biomed Inform.* **2014**, 47, 1–10.
 34. K. Liu, F. Li, L. Liu, Y. Han. Implementation of a kernel-based Chinese relation extraction system. *Jisuanji Yanjiu yu Fazhan(Computer Res Dev)*. **2007**, 44, 1406–1411.
 35. J. Zech, M. Pain, J. Titano, M. Badgeley et al. Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*. **2018**, 287(2), 570–80.