A Novel Pipeline for Outcome Prediction in Ovarian Cancer Using CT Radiology Reports

Sunantha Guruswamy
Assam Don Bosco University, Sonapur, Tepesia,Assam- 782402, India
Bobby Sharma
Assam Don Bosco University, Sonapur, Tepesia, Assam- 782402, India
Nilesh Sable
Tata Memorial Centre, Parel (E), Mumbai-400012, India
Satishkumar Chavan
Don Bosco Institute of Technology, Kurla (W), Mumbai-400070, India

Abstract
Problem: Electronic health careradiology reports contain diagnostic information, treatment details and drug dosage. Referring these reports on every follow up during the tenure of the treatment becomes tedious and unmanageable because of varying oncologists. Processing these textual radiology reports is challenging due to data imbalance, varying number of successive reports, changing format of the reports and differed flow of details of the organ. **Approach:** In this paper, a novel pipeline is proposed for structuring of report details, feature extraction & selection, and prediction of multilevel label. The research work is achieved using a proposed data wrangling algorithm for generating a structured dataframe (SDF). ML techniques are used for feature extraction & selection using TF-IDF of n-gram & TF-IDF of CBOW. Ahybrid transformer based deep learning (DL) technique is preferred for outcome prediction of ovarian cancer reports. **Findings:** The proposed pipeline is analysed on retrospective 984 CT radiology reports of 240 subjects during 2018-2020 treated at Tata Memorial Hospital, Mumbai. The algorithm achieved 100% information extraction followed by 489 unique features selection. The proposed hybrid transformer classifier method provided an accuracy of 96% and F1 score as 94%. **Conclusion:** The proposed hybrid transformer model has elevated the F1 score by6% to 12% when compared with state-of-the-artdeep learning or transformer methods.

Keyword: Ovarian cancer, Computed tomography, Radiology report, Transformer model, Bidirectional encoder representations from transformers, Long short-term memory networks, Natural language processing, Machine learning, Deep learning, structured dataframe.

1. Introduction

Ovarian cancer(OC)is the deadliest gynaecological cancer among women with an overall 5 years survival rate which is below 50% due to its asymptotic nature, diagnosis at advance stage, and high recurrence ratein 70% of the cases after standard therapies [1]. OCsare heterogeneous cancers where each sub type possesses a varied morphology and biology behaviour. Most ovarian cancers are carcinomas (epithelial origin) which predominantly fall into five histological subtypes: high grade serous, low grade serous, clear cell, endo-meteoroid and mucinous. Non epithelial ovarian cancer is much less common and includes germ cell, sex cord stromal and mesenchymal tumours. High grade serous carcinoma is the most common form of ovarian cancer accounting for approximately 70% of all cases. OCdoes not show any symptoms in it last stage, it is a difficult disease to detect[2].

In clinical practice, radiology reports (RR) in Electronic Health Care Reports (EHRs) format contain a lot of critical information and meta-data about patients (Fig. 1). The report is mostly available in free style in unstructured as well as sometime semi-structured format. In this research, the textis retrieved from 984

ovarian CT scan RR using natural language processing and machine learning modelled pipeline. Inthis paper text processing pipeline for various methods for information extraction, post structuring, feature selection (sentences) and classification of RR into multi-level-label for any domain type of cancer is feasible. The developed system will aid as an allied support radiologist and oncologist to provide a better health care support system in a short duration of time.

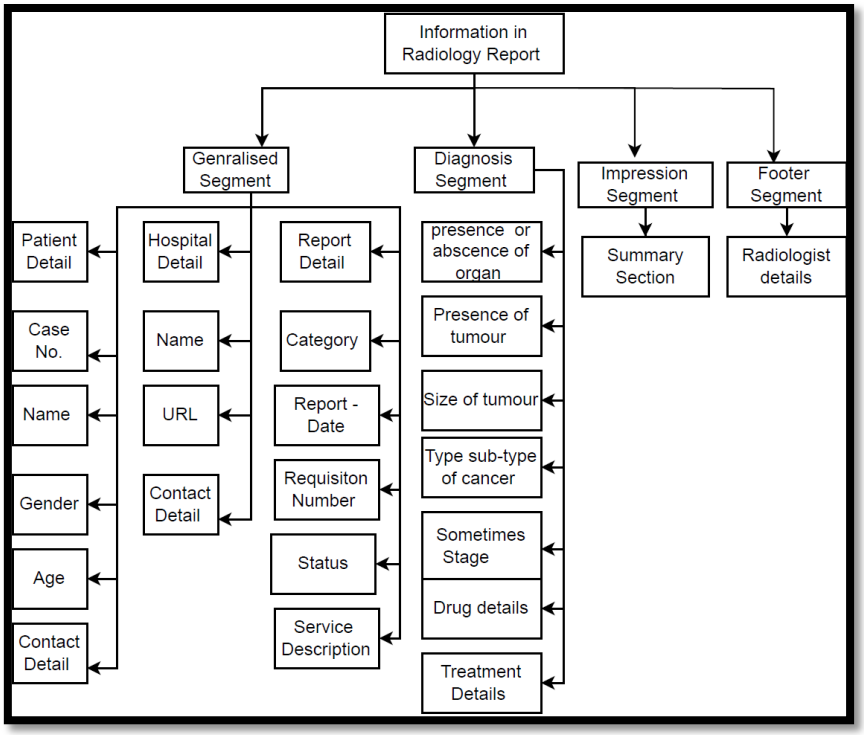

Fig. 1: Hierarchical form of critical information present in radiology report

Generic information extraction (IE) is achieved using tailored pattern matching approaches for structured format generation. In post-structuring, the structured format is cleaned and restructured to derive new features multi-level-labels, treatment progress, disease free survival rate that are beneficial in supporting classification of RRs of any domain of cancer. The objective of this research is to extract all information from all successive reports of a patient as a record into a structured format. This enables easy access to records during follow up without missing any investigation suggested in earlier reports.

The contributions of this research work are as follows:
i. A conditional rule-based pattern matching algorithm is proposed that retrieves 100% of the information from pdf format of EHR viz. patient detail, hospital detail, report transaction detail, diagnosis detail, drug detail, impression, and radiologist details.
ii. Hybrid feature extraction and selection methods extract and select important clinical annotations useful in classification of OC.
iii.Hybrid classification model consisting of Bidirectional Encoder Representations from Transformers (BERT) with Long Short-Term Memory Networks (LSTM) layers outperforms state-of-the-art classical models.

2. Background Study

Natural Language Processing (NLP) algorithms have a set of information extraction (IE) techniques like data mining, data wrangling, facts table, knowledge engine and data structure, to convert the CT scan RR (EHR/EMR) into structured data-frame (SDF) [3]. The state of the art NLP models [4-7] used inRR[8]along with ML [9-10]or deep learning (DL) [11-13]for extracting diagnosis text information only

from small section of the RR. The extracted information contains the size of tumour [14-17]and stage of the cancer [17-18] for only one domain of cancer. However, the challenging part is to extract the diagnosis details written in different languages [19-20]. Different text extraction techniques were applied to extract text from the diagnosis section namely Named Entity Recognition (NER) [21-22], semantic dictionary or word embedding (WE) techniques were used [23] and rule based feature extraction (RBFE) [24]. Few researchers were discussing multi-level-labelling[25] and organ specific multi-institutional cancer [16, 26-28] where our model provided better performance than these models.

Michael et al. [19] used appropriate models through NER and relation extraction through stratified sampling on images. The work was an adaptive fine tune method applied on imaging reports without any need to train the model on a downstream task. Approaches to use semi-automated process with GRU, for token level prediction of anchor entities and complete facts by using skip gram of vector with 300 dimension, to encode surrounding context for multiclass, multi-labelling[29]. Chng et al. [11] introduced a recommendation system with two subsystems namely knowledge base, contained facts and rules, rule based labellers, a search mechanism, which did not require heavy computation. The process of labelling chest x-ray radiology report by using transformer models is presented in [25]. Shreyasi et al. [30] suggested method of text extraction from reports based on top level heading converted into semi structured XML format. Yetisgenet al. [31] presented a text processing pipeline to automatically identify clinically important recommendation sentences and usedsupervised ML for text classification. Putra et al. [32] proved that feature extraction in reduces dimension is by word2vec. Sahdev et al. [20]introduced word embedding method, which constructs a global word-word co-occurrence matrix and utilizes matrix factorization to generate a multi-layer data representation.

In the last few years, various classification models were utilized to extract content based and image based CT scan RR diagnosis details in various domains of cancer using ML [33-36]and transformer based models [11]. Bendersky et al. [36] used logistic regression classifier for binary and multiclass classification on chest x-ray reports. Khosravi et al. [38] use radiology fusion with pathology reports and classified images into multiclass labels as benign, malignant, for prostate cancer. The work in [35] suggested a technique to classify cerebral tumour using SVM with a linear kernel into three categories by normalizing the MR images. Zhou et al. [33] proposed dynamiclanguage model classifiers (DLM) with naive bayes(NB), an automated classification model that gave an average accuracy. Multi-label classification of infectious disease like hepatitis, hand foot and mouth disease is achieved using deep learning model [39]. Putelli et al. [37] used DL classification model on chest CT in which it predicted the three classes between neoplastic, uncertain and non-neoplastic.

3. Data

With the approval from the ethical and legal committee, we obtained RRs from Tata Memorial Centre, Mumbai, India, for enhanced investigation of CT scan RRs for developing a hybrid transformer classifier model using BERT with LSTM that uses a novel IE method for extracting information from ovarian cancer CT scan radiology text report. The incident rate of subtype of OC falls under total 8 types and subtypes of OC which are serous(59.16%), stromal(08.75%), endometroid(04.56%), mucinous(05.00%), embryonal(00.01%), chorio-carcinoma(00.23%), teratoma(06.25%) and yolk sac(13.33%).

Following are the data collection criteria to limit and reduce any inconsistency that can improve the performance of the model than any other prevailing proven researches in this area.

i. Reports from 2018 to 2020 for all age group.
ii. Ovarian cancer CT scanRRs that are in PDF format.
iii. Reports of subjects, who have undergone only chemotherapy treatments.
iv. Cases with at least two consecutive scan reports.

v. Baseline(that does not contain treatment, stage, type of cancer), CT scan reports occupy 30% of the total data gathered that can be used for outcome prediction.
vi. Disease recurrence cases were excluded.

Table 1: Analysis of the radiology report gathered.

Parameter	Before 2018	2018 - 2019	After 2019
Total number of Bifurcations	2	3	4
Bifurcation details 1. General Segment 2. Diagnosis detail 3. Radiologist detail 4. Impression Segment	 Yes Yes	 Yes Yes Yes	 Yes Yes Yes Yes
Diagnosed organ details	Jumbled up	Unordered	Ordered

After 2019, the reports had 4 bifurcations that included additional section as 'impression' along with earlier 3 sections namely generalized section, diagnosis details, and radiologist details as shown in Table 1.

The following are the issues are faced while analysing the data
i. The number of segments and the order of sections within the segments are different across the report as the format has changed over the years (Table 1).
ii. Patient's details were not having definite format and length.
iii. The order of the disease diagnosis of each organ varied from report to report.
iv. Inter-observer variability in clinical vocabulary used in the reports.
v. Reports prior to 2019 did not have impression section.
vi. Consecutive reports may or may not contain the type, sub-type and stage of cancer.

The baseline reports are first scan reports of first hand report of the patient that do not contain the details like status of the treatment, stage, type or sub-type OC. The algorithm is applied and tested on both the types of reports, namely baseline report and the diagnosis successive reports.

4. Model

The steps of the proposed pipeline (Fig. 2) are as follows:
i. Conversion of RRs into intermediate files to recognize text from text files or image file format.
ii. Identification of segments and recognition text from the intermediate file.
iii. Perform conditional rule-based pattern search tailored approach to classify the extracted text and to store the classified data into a structured data frame.
iv. Post-processing of structure dataframe (SDF) improves theperformance of pipeline by creating, restructuring and deriving new feature namely multi-level-label, treatment status and treatment duration from successive reports of patients.
v. Perform feature engineering techniques to select features (TF-IDF of n-grams) from content field for DL methods and to create BERT tokens and tensor tokens for transformer models that has clinical importance and label encoding which can increase the accuracy.
vi. Predictive classification using proposed hybrid transformer method comprising of BERT with one layer of LSTM and comparison of performance with classical transformer models and classic DL models.

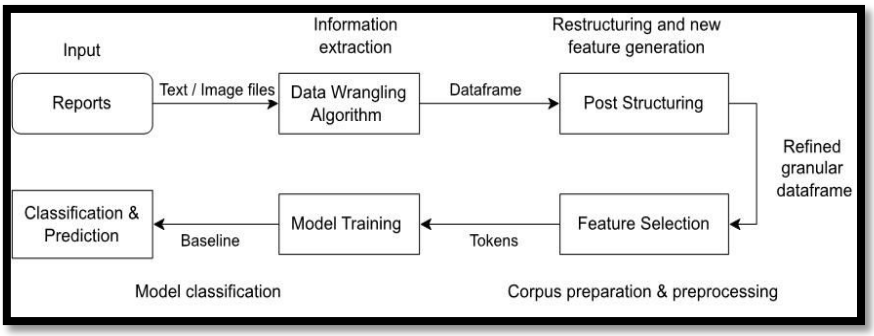

Fig.2: Proposed Model

4.1 Structured Text Extraction and Post-Structure Modelling

4.1.1 Text extraction

The first step of the pipeline begins with a preliminary conversion of the reports in PDF to intermediate file format. Various patterns are searched and the text is extracted using a proposed conditional rule based pattern search algorithm (CRPSA) as shown in Fig. 3. It consists of two methods:

Method 1: uses PyPDF2, Python Image Library (PIL) or pillow library, regular expression(re) and date function in python. PIL helps in converting cancer reports into bunch image of image files. Depending on the number of pages of the report, the number of images generated will also vary. So, this conversion will happen for all the records in the database. These bunch of image files became navigable and the string in these files will be read and recognized with the help of poppler and python-tesseract library using RPSA.

Method 2: uses python libraries namely openpyxl, re and date function in python. The pdfplumber is used to recognize the string in data and segregate various parts ofreport. The patterns of strings are searched and extracted implementing conditional rule based pattern search algorithm (CRPSA) toread, write and manipulation in the excel file which is much easier using openpyxl.

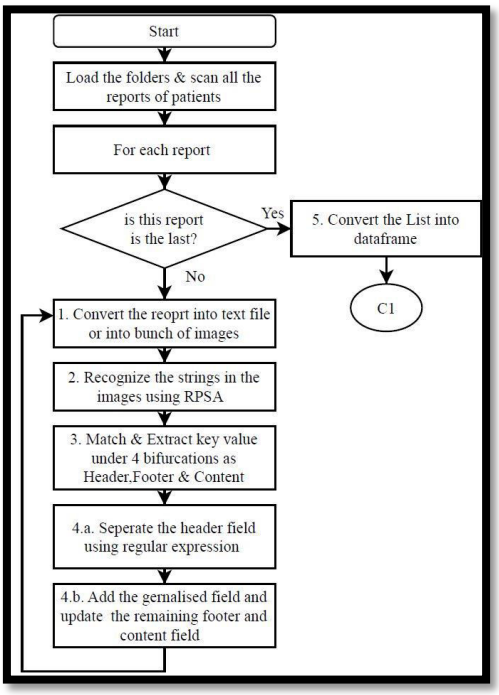

Fig.3: Text extraction using CRPSA - part 1

A data frame is generated with 3 columns namely header, footer, and content form extracted complex text. The header comprises of the patient detail and report detail. The content field contains the diagnosis details and impression. The footer contains radiologist details. This resultant standard accessible data frame generated by the text extraction methods has advantages like reduce the storage space of EHR, reduces the access time and improves the usage efficiency.

4.1.2 Post-Structure Modelling

Post-structuring modelling presents a second part of the CRPSA topic modelling algorithm 1. to clean and prepare inter record structuring by filtering or removing redundant records. 2. It also enriches the fields by transforming the pivot of successive reports (contents of reports in multiple rows into multiple columns of single row),3. Aggregating (content field) the records and deriving new generic field. New derived features are like multi-level-label, treatment status and treatment duration as in Fig. 3 and Fig. 4.

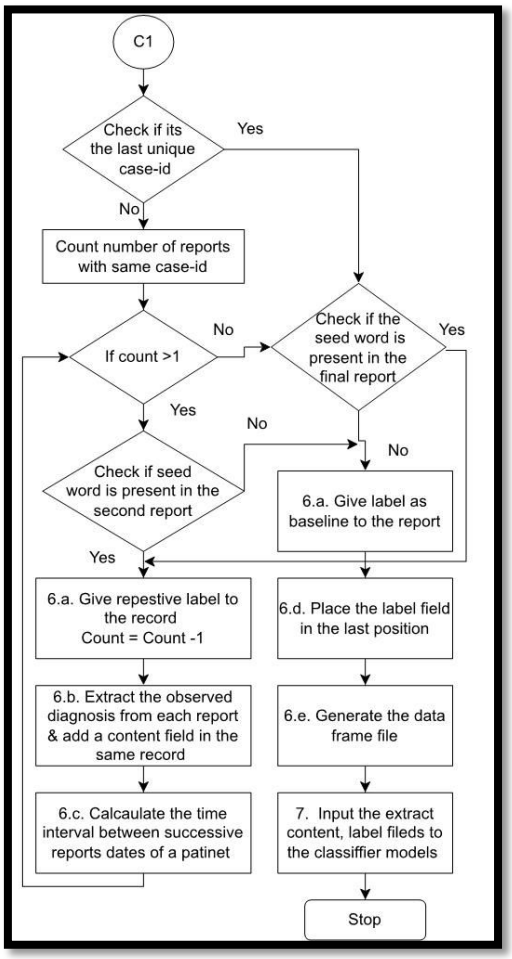

Fig. 4:Post-structuring and label extraction using CRPSA - part 2

4.2 Feature selection and preprocessing Method

Hybrid feature extraction method involves TF-IDF of n-gram by topic modelling which consists of sequence of steps to do standard pre-processing in generating token of words and feature vector using platforms viz.nltk, gensim and scala. Hybrid feature extraction has traditional NLP pre-processing steps to generate countvectorization of TF-IDF for DL models and steps for feature cleaning, token or tensor generation for transformer models.

4.2.1Pre-processing

In pre-processing, the content field of the generated structured file is converted into clinical vocabulary which undergoes dimension reduction. The steps are listed as follows.

i. Load the generalized structured data frame.
ii. Conversion of all the corpus data into lower case.
iii. Perform word tokenization or sentence tokenization.
iv.Remove all the stop words and punctuation.
v. Perform lemmatization to normalize the lexicons.

4.2.2 Feature selection

The clinical vocabularies or features from the preprocessed data are generated using NLP technique through a hybrid approach of Term Frequency-Inverse Document Frequency (TF-IDF) of n-gram.

The features are extracted using the following steps for the DL Models.

i. Generate n-grams of words that represent the more frequent tokens.
ii. Generate suitable feature vector of higher n-grams.
iii. Calculate the TF-IDF for n-gram that are above the threshold (threshold=0.5).

It enables generation of highly important clinical vocabularies from the content field of the structured data frame. These TF-IDF feature vectors are generated by replacing the vocabularies (word or sentence tokens) with their importance score and ready to fit the classification DL models.

The features are extracted using the following steps for the transformer models.

i. Generate the tokens using the content (diagnosis detail) column.
ii. Convert the token into tensor tokens which enables the tokens in tensor form ready to be fitted into transformer models.

4.3 Transfer learning based proposed hybrid transformer classifier model

Transfer Models (TF) helps to shorten the training time with fresh data and yields better results by using pre-trained models with already-fine-tuned weights and a big corpus of data. TF use the architecture of encoder-decoders. Many layers of attention layers are incorporated into the architecture of the transformer. The feed forward linear layer, attention layer, and self-attention model are the three main parts of the transformer model. Every word in the sentence has an attention vector produced by the self-attention model. Every word is parallelly mapped to every other word before being sent to the encoder and decoder blocks. These blocks transfer each attention vector to the feed forward linear layer and relate each word vector to the other. The probability distribution produced by feed forward linear layers is greater dimensional and produced more quick results. The encoder models used for these NLP tasks are ALBERTa, BERT, DistilBERT and ROBERTa and decoder models used are GPT and XLnet.

Table 2: Hyper parameters and training details for transformer models

Parameters	**Proposed, BERT, DistilBERT, RoBERT**	**GPT, XLNet**
Max Sequence	944	128
Training Epochs	10	10
Batch Size	16	16
Training Time (CPU minutes)	5	5

The transformer models were tunedusing hyper parameter namely maximum sequence, batch size, training time and count of tokens with a pretrained number of training time during the training phase are provided in Table 2.

5. Results

The tailored IE approaches have helped the classifier model to achieve a good accuracy.The proposed methodology involves three stages of experimentation as text extraction & post-structuring, hybrid feature extraction, and transfer learning based proposed hybrid classifiers model.

5.1 Text Extraction and Post-Structure

In the first step to convert each pdf file into an intermediate readable file format, like text or bunch of image files, using either of the text extraction approaches.During the second step, the intermediate files are retrieved and the text data or strings are recognized form the different segment using CRPSA as presented in Fig. 3 and Fig. 4 and the results and information are stored in a list as key value pair. Then the recognized string is inserted to SDF by creating and updating dataframe with key as the heading of the field and string value as value. The generalized details, radiologist details and the diagnosis details are captured into the header, footer, and middle section, respectively as shown in Fig. 5. Resultant dataframe contains 984 reports of patients as 984 record of the dataframe.

	Header	Footer	Content
0	CASENO: = Requisition No Name Mrs. Sex/Age : F/ 31 Years DMG: DMG-GYNEC ONCOLOGY Service Desc CT Thorax & Abdomen & Pelvis Category/Status : C/ Out Patient Rean Date : 08-11-2017	Dr.SAYED FAISAL Dr.ARPITA A SAHU Registrar(Radio-Diagnosis) Sr.Registrar/Consultant (Radio-Diagnosis) 2	CT SCAN OF THE THORAX, ABDOMEN AND PELVIS DATED 16.11.2017: Contrast enhanced scan of the thorax, abdomen and pelvis has been performed on a MDCT scanner. Clinical profile: case of choriocarcinoma No previous CT is available for comparison. THORAX Two enhancing lesions nodules are seen in the right lower lobe and left lower lobe larger one measures 2.7x1.8 cm in right lower lobe. The rest of the lung parenchyma &
1	CASENO: I Requisition No. Name Mrs. Sex/Age : F/ 28 Years Category/Status : C/ Out Patient DMG : DMG - GYNEC ONCOLOGY Service Desc CT Thorax & Abdomen & Pelvis Rean Date : 10-12-2019	Dr. VINEETH K.M. Dr.AKSHAY D BAHETI Registrar(Radio-Diagnosis) Sr.Registrar/Cc 3 RADIOLOGY	CT SCAN OF THORAX, ABDOMEN AND PELVIS DATED 10.12.2019: Contrast enhanced CT scan of the thorax, abdomen and pelvis has been performed on a MDCT scanner from root of neck till ischial tuberosities. This 1s a case of choriocarcinoma, post TAH and RSO, post chemotherapy (LD- 16.11.2019). No previous available for comparison. THORAX: Multiple (>20)
2	CASENO: J Requisition No. Name Mrs. Sex/Age : F/ 28 Years Category/Status : C/ Out Patient DMG : DMG - GYNEC ONCOLOGY Service Desc CT Brain Plain and Contrast Rean Date : 10-12-2019 Provisional Diagnosis Final Report Report Date: 11-12-2019	Dr. VINEETH K.M. Dr.AKSHAY D BAHETI Registrar(Radio-Diagnosis) Sr.Registrar/Cc 2 RADIOLOGY	CECT BRAIN DATED 10.12.2019 Plain and contrast enhanced scan of the brain has been performed on an MDCT scanner. This 1s a case of choriocarcinoma, post TAH and RSO, post chemotherapy (LD- 16.11.2019). Scan done to look for metastases No previous available for comparison. The cerebral hemispheres appear normal with no obvious focal or diffuse lesion. The

Fig.5: Result of PDF file converted to structured data frame

During the third step, patterns are searched by extracting the header field, using second part of the CRPSA which searched for keywords like case-id, name, report-date, requisition-number, category, requisition date, age and gender. These key value pairs are stored as individual new fields, and the remaining fields of the earlier dataframe are appended to its end. The resultant new dataframe is shown in Fig. 6.

CASENO	Requisition Number	Name	Age	Gender	Service Description	Requisition Date	Category	Status	DMG	Report Date
C1		P1	31	F	T Thorax & Abdomen &	08-11-2017	C	Out Patien	DMG-GYNE	08-01-2018
C2		P2	28	F	T Thorax & Abdomen &	10-12-2019	C	Out Patien	DMG - GYN	11-12-2019
C3		P3	28	F	T Brain Plain and Contr	10-12-2019	C	Out Patien	DMG - GYN	11-12-2019
C4		P4	29	F	T Thorax & Abdomen &	19-01-2021	NC	Out Patien	DMG-GYNE	22-01-2021
C5		P5	11	F	T Thorax & Abdomen &	24-01-2020	C	Out Patien	DMG - PAE	30-01-2020
C6		P6	11	F	T Thorax & Abdomen &	24-01-2020	C	Out Patien	DMG - PAE	30-01-2020
C7		P7	32	F	T Thorax	28-09-2021	C	Out Patien	DMG - GYN	30-09-2021
C8		P8	8	F	T Thorax & Abdomen &	08-06-2021	B	Out Patien	DMG - PAE	09-06-2021
C9		P9	8	F	T Thorax & Abdomen &	15-06-2021	B	Out Patien	DMG - PAE	13-09-2021
C10		P10	43	F	T Thorax & Abdomen &	22-01-2016	B	Out Patien	DMG - GYN	22-02-2016
C11		P11	20	F	T Abdomen & Pelvis	16-03-2017	C	Out Patien	DMG - GYN	31-03-2017
C12		P12	82	F	T Abdomen & Pelvis	17-05-2017	C	Out Patien	DMG - GYN	31-05-2017
C13		P13	38	F	T Abdomen & Pelvis	08-11-2017	B	Out Patien	DMG - GYN	10-11-2017
C14		P14	25	F	T Abdomen & Pelvis	28-02-2018	C	Out Patien	DMG - GYN	27-03-2018
C15		P15	39	F	T Abdomen & Pelvis	10-12-2018	C	Out Patien	DMG - GYN	13-12-2018

Fig.6: Segregation of the Header Column into individual fields of features

The fourth step is post-structure processing which removes redundant records, restructures the records and transforms the dataframe from 984 records into 240 records. Matching each unique case-id of the reports are extracted from the rows and only the content column and report date of the consecutive reports is

transformed and updated as new column. Thus, for each successive report of a patient, 2 columns are updated to the record. On the third run of the algorithm, the diagnosis details form content columns are extracted as tokens using NLP preprocessing steps and new facts are derived as multi-level-labels based on type and sub-type of OCusing CRPSA algorithm part 2. It results in fine granulated relational inter record fields. Fig. 7 shows the new SDF with multi-level-labels.

Label1	Label2	Treatment Status	Report Content	Report Conten	Report Date 3	Report Date 4	Duration 1	Duration 2	Duration 3	Total Duration
serous		stable	CT SCA	Final F	30-07-2020		132	167		299
serous		stable	Final Re	Final F	29-01-2019		152	90		242
baseline		baseline	Final Re					0	0	0
stromal		Stable	Final Re	Final F	06-09-201	09-04-20	187	168	580	935
serous		Regressior	Final Re	Final F	21-11-2020		286	195		481
	Sex Cord	Regressior	Final Re	Final F	24-08-2016		48	365		413
baseline		baseline	Report							0
serous		Stable	Final Re	Final F	08-07-2021		10	65		75
serous		progress	Final Re	Final F	18-02-2019		123	84		207

Fig.7: Snapshot of derived features: multi-level-label, treatment Status and disease free duration

5.2 Feature selection and preprocessing

The corpora of all documents are prepared from the content column of all the records. On this data, words tokenization and sentence tokenization are applied to extract the features. The results using nltk, gensim and Scala for both word and sentence tokenization is summarized in the Table 3.

Table 3. Summary of preprocessing results

Parameters	Gensim	Scala	Nltk
Word count before cleaning	69360	51314	175518
Word count after cleaning	489	653	885
Count of unique words	489	653	678
Top 5 words in TFIDF	seen	seen	Seen
	normal	normal	Normal
	unremarkable	right	Right
	right	left	Left
	pelvis	ct	ct
Count of tokenized Sentences	6412	6844	9763

The identified key features or clinical vocabulary for further data analysis help in dimension reduction process. It is done with help of the NLP and ML techniques of feature vector generation using of n-grams with TF-IDF. The results achieved through hybrid feature extraction are clinically correlated highly important tokens which are considered as features (Fig. 8). Further the features TD-IDF values are generated for n-gram method of 7 grams, but the sample in Fig. 8 shows TF-IDF score of 3 grams for feature selection.

```
#keep track of feature name and its corresponding score
        score_vals.append(round(score, 3))
        feature_vals.append(feature_names[idx])

arising right adnexal 0.211
solid cystic lesion 0.157
right adnexal region 0.149
lymph nodes size 0.121
nodes size criteria 0.118
seen bilateral apical 0.113
scattered lungs show 0.113
scan study available 0.113
sac tumour ovary 0.113
pre chemotherapy ct 0.113
posteriorly otherwise unremarkable 0.113
```

Fig. 8: Sample result of selected features with TF-IDF score

5.3Transfer Learning based proposed hybrid Classifiers model

The content field, and the label are selected form the data frame and are cleaned, preprocessed and tokens are generated. Further the countvectorization of TF-IDF was calculated for the tokens and later was split into train and test (75:25) sets. We used three layers for all DL models and trained and tested with imbalanced data set as well as balanced data set. The performance measure was calculated in terms of accuracy, sensitivity, precision and F1 score.

Table 4: DL model performance with imbalance data

Model	Precision	Recall	F1	Specificity	Accuracy
ANN	19	44	27	87.5	44
RNN	21	39	27	87.8	39
CNN	72	68	62	93.3	68
LSTM	19	43	26	87.5	43
BiLSTM	10	32	15	87.5	32

Table 5: DL model performance with balanced data

MODEL	Precision	Recall	F1-score	Support	Specificity	Accuracy
ANN	92	89	87	227	98	89.42
CNN	93	92	89	227	98.57	89.86
RNN	91	88	87	227	98.49	89.42
LSTM	38	12	47	80	87.5	61.25
BiLSTM	30	55	39	80	87.5	55

Among all DL models, CNN shows better performance over state-of-the-art DL models. CNN provides 89.42% accuracy, 98.57% specificity and 89.00% of F1 score in balanced data set. CNN performs better than LSTM as CNN processes sequence of words based on time series prediction. CNN provides better accuracy than LSTM with comparatively with less time. The performance evaluation for all the multi-level-label text classifiers of DL shown with balanced data(Table 5) and imbalance data (Table 4)

Multi-class classification using pre-trained attention models like BERT, RoBERTa, DistilBERT with fine-tuned layers is experimented. With one layer of LSTM classifier, the transfer learning architecture generates sequential optimized attention vector. The transformer model performance for balanced and imbalanced data set is presented in Table 6 and Table 7, respectively. The results show that the

performance of BERT with LSTM hybrid model for balanced dataset is increased by 18.5% over imbalanced dataset. The hybrid classifier model (BERT with LSTM) provides accuracy 96.00%, specificity 99% and F1 score of 94.00%. The graphical representation of accuracy and training loss with proposed BERT+LSTM Model is presented in Fig. 9.

Table 6: Transformer model performance with imbalance data

Model	Precision	Recall	F1-score	Specificity	Accuracy
BERT	20	32	24.62	87.5	29
RoBERT	72	66	68.86	97.25	82
DistillBERT	94	65	76.86	97.87	82
GPT	22	26	23.83	89.01	29
XLNet	58	48	52.53	94.23	60
Proposed	77	73	74.95	96.97	88

Table 7: Hybrid transformer model performance with balanced data

Model	Precision	Recall	F1-score	Specificity	Accuracy
BERT	33	34	24.62	89.31	49
RoBERT	89	89	68.86	90.25	86
DistillBERT	88	88	76.86	97.24	80
GPT	58	41	23.83	96.95	42
XLNet	70	89	52.53	97.11	72
Proposed	94	94	74.95	98.99	96

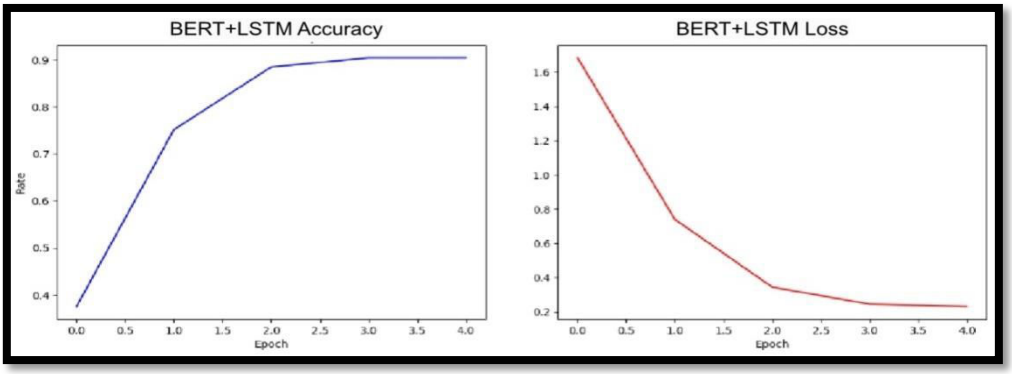

Fig. 9: Graphical representation of accuracy and loss of the proposed model

6. Discussion

6.1 Performance of different extraction models:

Given any CT scan RR in the PDF file format, it is converted into a structured data frame (Fig. 5). The proposedpipeline contains various stages. During the initial stage, python libraries are used to make the PDF file content readable by converting them into bunch of text files or into bunch of image files. The intermediate conversion of RR from PDF format into readable form is performed using method 1 and method 2 of proposed CRPSA IE algorithm. Method 1 converts pdf into text files and method 2 generates bunch of image files without any data leakage, noise, and inconsistencies. However, direct usage of PyPDF2 had issues in reading PDF files to recognize the some strings (alphabets) and the key value pairs get misaligned or aligned in different line thus inducing noise or error. The proposed algorithm part 1 is

able to recognize all the strings and values appropriately and generates data frame successfully. The generated dataframe file contains records of the total number of reports as shown in Fig. 5. The resultant SDF contains 984 rows where each consecutive reports of a patient form the row in the dataframe.

6.2 Performance of different feature extraction models:

The second part of algorithm (post structuring) helps to clean and to purge empty and redundant records. CRPSA algorithmderives features from content field by preparing a corpus and generating multilevellabels.NLP preprocessing steps involved generating tokens of sentences and words are shown in Fig. 8. As a result, the feature extraction is successful even though there are no fixed spaces and positions for the strings in the various segments of the report. To generate highly important features of cancer domain, tokens are generated by n-gram and TF-IDF through sklearn.

6.3 Performance of different classification models:

Automated classification of OC CT scan RR using DL model does not perform as expected due to insufficient, imbalance amount of data under each sub-classification level of ovarian classes. Out of all the DL models, CNN model has achieved an accuracy of 89.42% and F1 score of 89%. There is an increase of F1 score from 16% to36% when data is made balanced by augmentation. As seen in Table 6 and Table 7, hybrid transformer model (HTM) (BERT with LSTM) performed better than other transformer models with an accuracy of 96% and F1 score of 94%.

Lung carcinoma classification using PET-CT scan RR using BiLSTM model in [40] outperformed than XG Boost with an overall F1 score of 94.0%. It is recommended that pre-trained models can increase the performance of classification models when the data is less as suggestions by [11]. The classification in [37] on chest tomography has gained an accuracy of 80.0% to classify neoplastic vs non neoplastic. NER (Pre-existing library for feature selection) with BERT model provided 86.97% of F1 score [20]. As compared to the state-of-the-art techniques, the proposed algorithm with tailored IE has helped CNN (DL model) to achieve F1 score and accuracy of 89.42% and 89.00%, respectively. HTM has achieved 96% accuracy.

The limitation of proposed work is that entire pipeline approach is customized for only text extraction from pdf file format report. The tailored algorithm used for 100% IE using CRPSA that generated SDF. After post-structuring, only diagnosis details form content field are used to identify labels that are recommended in any one of the consecutive reports of patients. The data gathered is imbalanced in 8 types and sub-types due to one disease (serous) prevalence's over the others. BERT when combined with a layer of LSTM, a DL models, the performance of the classifier model increased from 6% to 12%.

7. Conclusion

Automated multi-level-labelled classification of content-based CT scan RR using pipelined NLP allows extracting and deriving multi-level-labels of ovarian cancer or it can be adapted for other domains of cancer. The proposed algorithm presents a novel IE technique, hybrid feature selection and hybrid transformer classifier (BERT + LSTM) model for outcome prediction. This work archives text extraction (patient, hospital, report, diagnosis, impression, and footer details) with 100% consistency. A HTMis evaluated on balanced and imbalanced dataset which achieved accuracy and F1 score of 96.00% and 94.00%, respectively. Due to the functioning of BERT for parallel generation of sequence of statements or facts and LSTM for classification of text embedded in proposed algorithm, it outperforms over the state-of-the-art techniques. The proposed algorithm gets trained successfully with lesser amount of data.

The presented algorithm helps in minimizing secondary access for record through generated SDF. It also supports in secondary usage like development of recommendation system or expert system or report generation with improved accuracy rate. Automatic classification can assist radiologists in critical decision

making based on the automatic multi-level labelling generated. A future work will be in developing a recommendation system or an expert system for suggesting critical patient management.

References

1. Lee, Y. T., Tan, Y. J., & Oon, and C. E. (2018). Molecular targeted therapy: Treating cancer with specificity. *European journal of pharmacology*, 834:188-196.
2. Breen, J., Allen, K., Zucker, K., Adusumilli, P., Scarsbrook, A., Hall, G. and Ravikumar, N. (2023). Artificial intelligence in ovarian cancer histopathology: a systematic review. *NPJ Precision Oncology*, *7*(1): 83.
3. Datta, S., Bernstam, E. V., & Roberts, and K. (2019). A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of biomedical informatics*, 100:103301.
4. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., and Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14-29.
5. Mithun-Nair, S., Jha, A., Rangarajan, V., Wee, L., and Dekker, A. (2021). Natural language processing in radiology reports.8:461–472.
6. Pons, E., Braun, L. M., Hunink, M. M., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329-343.
7. M Kumbhakarnaa, V., Kulkarni, S., and D Dhawaleb, A. (2020). Clinical text engineering using natural language processing tools in healthcare domain: a systematic review. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.
8. Nandhakumar, Nidhin & Sherkat, Ehsan & Milios, Evangelos & Gu, Hong & Butler, Michael. (2017). Clinically Significant Information Extraction from Radiology Reports. 153-162.
9. Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., and Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
10. S. Jewel. (2019). Natural language-based machine learning models for information extraction from radiology reports –survey. International Journal of Science and Research (IJSR), 8:273–277.
11. Chng, S. Y., Tern, P. J., Kan, M. R., and Cheng, L. T. (2023). Automated labelling of radiology reports using natural language processing: Comparison of traditional and newer methods. *Health Care Science*, *2*(2):120-128.
12. Sugimoto, K., Takeda, T., Oh, J. H., Wada, S., Konishi, S., Yamahata, A., and Matsumura, Y. (2021). Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.
13. Martín-Caro García-Largo, M. Á., and Segura-Bedmar, I. (2021). Extracting information from radiology reports by Natural Language Processing and Deep Learning. Ceur Workshop Proceedings.
14. Yamashita, R., Bird, K., Cheung, P. Y. C., Decker, J. H., Flory, M. N., Goff, D., and Desser, T. S. (2021). Automated identification and measurement extraction of pancreatic cystic lesions from free-text radiology reports using natural language processing. *Radiology: Artificial Intelligence*, *4*(2):e210092.
15. Bozkurt, S., Alkim, E., Banerjee, I., and Rubin, D. L. (2019). Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *Journal of digital imaging*, *32*: 544-553.
16. Saha, A., Burns, L., & Kulkarni, and A. M. (2023). A scoping review of natural language processing of radiology reports in breast cancer. *Frontiers in Oncology*, *13*:1160167.
17. Senders, J. T., Karhade, A. V., Cote, D. J., Mehrtash, A., Lamba, N., DiRisio, A., and Arnaout, O. (2019). Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. *JCO clinical cancer informatics*, 3:1-9.
18. Sahdev, A. (2016). CT in ovarian cancer staging: how to review and report with emphasis on abdominal and pelvic disease for surgical planning. *Cancer Imaging*, *16*(1):1-9.

19. Jantscher, M., Gunzer, F., Kern, R., Hassler, E., Tschauner, S., and Reishofer, G. (2023). Information extraction from German radiological reports for general clinical text and language understanding. *Scientific Reports*, 13(1): 2353.
20. Liu, H., Xu, Y., Zhang, Z., Wang, N., Huang, Y., Hu, Y., and Chen, H. (2020). A natural language processing pipeline of chinese free-text radiology reports for liver cancer diagnosis. *Ieee Access*, 8:159110-159119.
21. Steinkamp, J., Chambers, C., Lalevic, D., & Cook, T. (2021). Automatic fully-contextualized recommendation extraction from radiology reports. *Journal of Digital Imaging*, 34:374-384.
22. Obuchowski, A., Klaudel, B., and Jasik, P. (2023). Information Extraction from Polish Radiology Reports using Language Models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing (SlavicNLP 2023)*, 113-122.
23. Banerjee, I., Chen, M. C., Lungren, M. P., and Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of biomedical informatics*, 77:11-20.
24. Sykes, D., Grivas, A., Grover, C., Tobin, R., Sudlow, C., Whiteley, W., and Alex, B. (2021). Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, 27(2): 203-224.
25. Schrempf, P., Watson, H., Park, E., Pajak, M., MacKinnon, H., Muir, K. W., and O'Neil, A. Q. (2021). Templated text synthesis for expert-guided multi-label extraction from radiology reports. *Machine Learning and Knowledge Extraction*, 3(2): 299-317.
26. Hassanpour, S., and Langlotz, C. P. (2016). Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29-39.
27. Lazic, I., Jakovljevic, N., Boban, J., Nosek, I., and Loncar-Turukalo, T. (2022). Information extraction from clinical records: An example for breast cancer. In *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON),* 942-947, IEEE.
28. Nobel, J. M., Puts, S., Weiss, J., Aerts, H. J., Mak, R. H., Robben, S. G., andDekker, A. L. (2021). T-staging pulmonary oncology from radiological reports using natural language processing: translating into a multi-language setting. *Insights into Imaging*, 12(1):77.
29. Steinkamp, J. M., Chambers, C., Lalevic, D., Zafar, H. M., and Cook, T. S. (2019). Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32:554-564.
30. Pathak, S., van Rossen, J., Vijlbrief, O., Geerdink, J., Seifert, C., & van Keulen, M. (2019). Post-structuring radiology reports of breast cancer patients for clinical quality assurance. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):1883-1894.
31. Yetisgen-Yildiz, M., Gunn, M. L., Xia, F., & Payne, and T. H. (2013). A text processing pipeline to extract recommendations from radiology reports. *Journal of biomedical informatics*, 46(2): 354-362.
32. Putra, S. J., Gunawan, M. N., and Hidayat, A. A. (2022). Feature Engineering with Word2vec on Text Classification Using The K-Nearest Neighbor Algorithm. In *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, 1-6, IEEE.
33. Zhou, Y., Amundson, P. K., Yu, F., Kessler, M. M., Benzinger, T. L., and Wippold, F. J. (2014). Automated classification of radiology reports to facilitate retrospective study in radiology. *Journal of digital imaging*, 27:730-736.
34. Nguyen, D. H., & Patrick, and J. D. (2014). Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, 21(5):893-901.
35. Ayadi, W., Charfi, I., Elhamzi, W., and Atri, M. (2022). Brain tumor classification based on hybrid approach. *The Visual Computer*, 38(1): 107-117.
36. Bendersky, M., Wu, J., and Syeda-Mahmood, T. (2018). Classification of radiology reports by modality and anatomy: A comparative study. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 1457-1464, IEEE

37. Putelli, L., Gerevini, A. E., Lavelli, A., Olivato, M., and Serina, I. (2020). Deep learning for classification of radiology reports with a hierarchical schema. *Procedia Computer Science*, 176: 349-359.
38. Khosravi, P., Lysandrou, M., Eljalby, M., Li, Q., Kazemi, E., Zisimopoulos, P., and Hajirasouliha, I. (2021). A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion. *Journal of Magnetic Resonance Imaging*, *54*(2):462-471.
39. Wang, M., Wei, Z., Jia, M., Chen, L., and Ji, H. (2022). Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. *BMC medical informatics and decision making*, 22(1): 1-13.