ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse25952020.pdf

https://doi.org/10.30534/ijatcse/2020/25952020

Advanced Information Extraction System based on Unstructured Mammogram Reports

Ghadah Aldabbagh

King Abdulaziz University, Department of Computer Science, Jeddah, Saudi Arabia

Email: galdabbagh@kau.edu.sa

ABSTRACT

In this paper, the aim is to develop a solution to help the breast cancer initiatives in achieving their goal in Saudi Arabia by building a system using a Natural Language processor (NLP) for processing mammogram reports to organize and structure the collected information and identify particular abnormalities. The proposed system can speed up the diagnosis process and save a patient's life. The system has been implemented by using a spaCy and NLTK library with Python programming language. The system obtains an accuracy of 90.9%, Precision 77.97 %, Recall 76.14 % and F1-score 74.4% with respect to the actual information stored in each frame. The proposed system can be extended in other radiology domains supporting large-scale data mining or in complex decision making frameworks.

Key words : Mammogram Reports, Information Extraction, Natural Language Processing.

I. INTRODUCTION

Breast cancer is the principal type of cancer causing fatalities in the female population of Saudi Arabia. [1] (Bray et al., 2018). As reported in [1] approximately 3,629 new breast cancer cases have been diagnosed in Saudi Arabia till 2018. Estimated number of the new cases of each type of cancer in Saudi Arabia at 2018 is given inFigure 1. Furthermore, the breast cancer incidents are estimated to expected to raise by 350% till 2025 [2]. With this rapidly increasing, there are many initiatives for the early diagnosis[3] because this can likely lead immediate treatment and to full recovery [4]. Unfortunately, these invitations could not limit the increased numbers of breast cancer incidence and more cases died as a result of late detection [5]. In this paper, we seek to address the challenges facing the breast cancer early detection invitations and find the appropriate solution that can enhance their role in the reduction of related fatalities.

In Saudi Arabia, there are many initiatives for the breast cancer treatment and early diagnosis [6]. Even so, the International Agency for Research on Cancer recorded around 899 out of 3629 cases that die as a result of late diagnosis in 2018 only, as shown in Figure 2[1]. This means that initiatives could not achieve their goal as it should be. Among the reasons for that failure, diagnosis in Saudi Arabia requires a double reading of the mammogram by two different radiologists which makes the diagnosis consumes a very long time [7]. In double reading procedure, the first radiologist writes a report describing the finding from the mammogram while the second radiologist reviews these finding and assess the results. The report writing by the first radiologist contains a lot of information in an unstructured format, which causes some difficulties for the second radiologist when evaluating the results. Furthermore, there is lacking in the radiologists in Saudi Arabia [8].

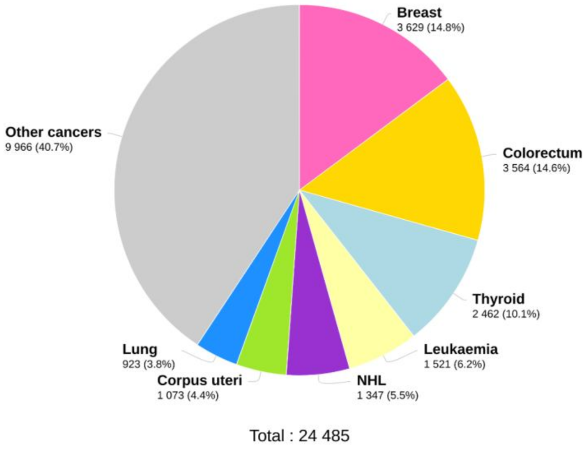


Figure 1: Estimated number of all cancers new case in 2018, Saudi Arabia [1]

This paper suggests a solution to help the breast cancer initiatives in achieving their goal in Saudi Arabia. We proposed to develop a system using Natural Language Processor (NLP) techniques aim to help radiologists in diagnosing by the mammogram double reading procedure. The proposed system gets the mammogram unstructured report as input, then extracts important information in "information frames" format as output. Thus, the radiologists can diagnose more patients in less time and efforts. As a result, the number of cases that are detected in late stages will be reduced. Target users are research centers and radiologist department in Ministry of National Guard Health (NGH) hospital. Specifically, the breast unit department that diagnoses breast cancer.

The related state-of-the-art methods are discussed in Section 2, followed by the algorithmic analysis and design of the proposed system in Section 3. The prototype and mobile app development are presented in Section 4, along with illustrative screenshots and functionalities. Section 5 illustrates the test results of the system and, finally, Section 6 concludes the paper.

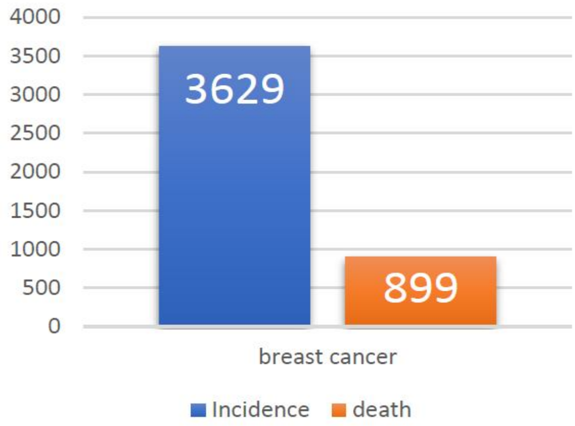


Figure 2: Estimated number of incident breast cancer cases and deaths among female in 2018, Saudi Arabia

II. RELATED STATE OF THE ART

The proposed work by [9] discussed the text formatting and storage of health-related information with no additional processing. This work proves that searching and manual review of such data is cumbersome, which agreed the need of using NLP software for organizing information elements contained in electronic health records. Key difficulties with current NLP methods are the necessity for synonyms or ontology dictionaries, which often available only in English. On the contrary, their observations are written in French. Therefore, their work involved creating a group oncologists for providing their expertise in developing novel solutions using NLP for enabling detailed analysis of medical information recorded in free text through the use of a synonym detection algorithm coupled with the construction of the search. A maximum of around 14k anonymized notes referring to around 9.6k breast cancer diagnoses was used to achieve extraction accuracy ranging from 70 to 96.8%.

The model proposed by [10] explores the possibility of an automated knowledge extraction method from Norwegian pathology records. A pathologist who analyzed tissue samples belonging to diagnosed cancer cases or to cases with large likelihood to have cancer. They provided a compilation of 40 pathology reports detailing breast cancer tissue samples. The document contained several test results, measurements and sample details used to establish methodologies for extracting information based on certain rules. In order to assess the quality of this model, its outputs were compared against those generated by expert individuals performing manually encoding of similar pathologies and twenty five experts who transferred data from manually written documents to databases using XML models. On average, the model result is 80 % accuracy, 98 % recall and 86 % F-score.

The researcher [11] focused on developing electronic medical records (EMR). EMR information can be broken down into three data types: structured, semi-structured, and unstructured. The unstructured type(e.g. radiology or pathology reports) holds many valuable data but lacks specific conceptual structures, and may also include errors in grammar, spelling, dialects, etc. This work proposed using certain algorithms, including a rule-based system, to identify the entity's rules from a related text, as well as suggested the rules that were applicable only to unique dataset collections. The resulting Fscore was 0.715 for correct matches and 0.824 for overlapping matches. Additionally, structural support vector machines (SSVM) were investigated and led to the algorithmic design incorporating exploiting the benefits of conditional random fields (CRFs) as well as SVMs, and also word representation feature sets. The proposed system of SSVM had the highest Fmeasure (close to 89%), outperforming the best system recorded by 0.6%, indicating that SSVM is has strong potential in medical NLP applications.

Additionally, the researchers [12] proposed the use of NLP for extracting mammogram observations. This system focused on extracting four finding types, related to asymmetry, architectural distortions, object mass and calcification via the use of methods based on dictionary look-up. The dataset used for this system contains around 93.7K mammogram report samples by Group Health in Seattle, Washington. The process model includes preprocessing step and rule-based NLP algorithm step. In addition, confidence flags have been created to indicate highly confident NLP reports and reports having likely errors. The system could successfully code 96–99 out of 100 random samples that manually abstracted. All findings achieved sensitivity, specificity, and negative predictive values above 0.92. However, there have been limitations on the NLP system, the NLP system was unable to differentiate between current and historical results.

The proposed model by [13] demonstrated the techniques of machine learning that can reliably interpret reports of pathology without manual effort. By training a model based on machine-learning on breast pathologies reports for extracting specific tumor characteristic features. The work-flow started by taking around 91,5K breast pathologies reports from three different hospitals. These reports exhibited marked differences in verbalizations and did not follow the same structure. All reports were anonymized by removing elements such as stripping date, number of medical record, and name of patient. The pathology report and the annotations for each pathology reports which annotated by trained professionals is input of the system. The system use boosting algorithm to classify pathology reports and this end with 20 separate categories of information, including types of atypia (structural abnormality in a cell) and specific tumor characteristics. The output of the system is a database with 20 columns (separate categories) and having 108,114 rows (records). The system achieved 90 % accuracy for correct analysis of all types of carcinoma and a typia for a particular patient.

Furthermore, the model proposed by [14] aimed to automate Breast Imaging-Reporting And Data System (BI-RADS) category types extraction from mammogram report samples. The model divided the extracting stage into two sub-stages: (a) annotating BI-RADS categorical values in the reports, (b) classifying of the laterality of the category value of the respective BI-RADS. The first task was done by developing a rule-based method to annotate all BI-RADS values related to risk assessment in the mammogram report. Each token throughout the text was characterized as “BI-RADS category value” or “not BI-

RADS category value", judging from the context information within the surrounding lines of the text. The second task involved developing automatic process for the classification of the resultant annotations as (i) Bilateral (ii) Nonspecific, (iii) Overall. (iv) Left, (v) Right. The problem formulation led to the definition a multiclass classifier problem. For this task, a comparison between models based on Support Vector Machines (SVMs) and Naïve Bayes (NB) was applied to determine best classifier. In the end, the SVM model achieved 0.9 for f-score highly than the NB model where is achieved 0.87 f-score. However, as supervised model there is a need to manually classification and annotation for all documents before working.

Also, the goal of the model that proposed by [15] was the performance assessment of NLP in the extraction of abnormal findings from free-style mammogram text and their conversion using structured formats that can be used in Clinical Decision Support Systems. This model looks over the reports to detect the term "IMRESSION:" and to extract the BI-RADS results that appear afterwards. When NLP was compared to manual review for mammogram reports, the results were as follows: 98% for precision, (98–100%) for recall, finally the accuracy was 98%.

The authors in [16] suggest extracting information automatically out of mammogram reports. The used approach first Explore all the relations in the text report. Following that, the unsupervised word2vec algorithm was used. Finally, the extracted relations and associated entities are linked to generate the information frames. Although the system successfully extracted from the reports more comprehensive information than the rule-based system, it has some limitations. First, there is a limited number of the 162 relations extracted in the research, because there were few reports available. Secondly, the approach does not link between different relations or extract complex relations. Finally, the approach used "Stanford POS tagger" for relation identifying and extracting, that is not fully accurate when parsing text reports related to radiology. The approach performed well in extracting information from mammogram reports, with 94% accuracy.

The work in[17] developed techniques for processing and extracting information from medical reports, converting it into a structured format. Unstructured reports have been converted to structured and different processing has been done using NLTK text processing software such as, case conversion, removal of special numerals and tokenization. This technique uses word2vec to represent word in vector space. Then K-means clustering algorithm was trained to cluster all the word. The extract feature used to malignant/benign cancer classification using boosting classifier. The system achieves 89% for both f-score and accuracy.

The proposed model by [18] analyzes the narrative form mammogram report in order to infer the BI-RADS final assessment categories. The proposed unsupervised approach combines embedding semantic terms with distributed semantic terms, leading to vectoral representations of mammogram reports. By training a machine learning (ML) model on (300,000) unannotated mammogram reports, it became possible to output contextual vectoral BI-RADS report representations. This resulted in a score of 87% for all of the precision, recall, and F1-score metrics.

Our system will involve an unsupervised approach to extract information automatically.

III. ALGORITHM DESIGN AND IMPLEMENTATION

In order to implement the automated information extraction model from unstructured reports, there is a need to build a machine learning model that learns from a dataset and predict the results for a particular task [19]. The proposed system uses unstructured text as a dataset. Moreover, the text data often contains a lot of information, which will make applying any machine learning algorithms are complex task. Therefore, it is extremely important to pre-process the text to prepare the data for machine learning methods in order to solve a variety of problems satisfactorily. Text preprocessing involves the sequential steps of sections segmentation, sentence splitting, tokenization, and Part Of Speech (POS) tagging.

This section will present the algorithm which takes the unstructured report as input and then produces a structured report (information frame). The algorithm divides into three steps: (a) extraction of relation: this step about extract the relation takes report text and the output will be a list of relation words, (b) clustering of relation: this step will take the output of the previous step as input and it is output will be a groups of clusters similar relations, and (c) associating relations phase: this step will be linking the extracted relations with the related entities and automatically entitle the frames according to the class to which the relation belongs.

To extract the relation from the reports, we used a spacy library because it offers many helpful and state-of-art algorithms and its performance is usually great as compared to NLTK. Clustering relation is the second phase during which similar relation types re clustered for providing an improved view of mutual relation types among entities. The cluster done by using Scikit-learn library. We use k-means and Mini Batch K-Mean cluster algorithms. The focus of Scikit-learn is to bring machine learning to non-specialists using a high-level general purpose language [20].

K-means is one of the most widely used clustering schemes, and have extensive application in clustering and analysis of information with good time performance [21]. The algorithm selects a different initial center for the same dataset and k value for each run of the algorithm, and then it follows an optimization procedure by choosing a different starting center. The determination of the initial cluster center affects the results. The relation between two objects is assessed via using their Euclidean Distance [22]:

$$d_{\text{Euclidean}}\,(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

where:
d(x,y)is the distance of data x to the center of the cluster y.

Mini Batch K-means is a clustering version of K-means with less complexity and less time consuming. For each iteration the concept of Mini Batch K-means is to use a small amount of training data. This speed up convergence and greatly shortens training time. This methodology is also used in other fields such as artificial neural networks to reduce back propagation algorithm training time [23].

After implementing the two algorithms, we compare between it in terms of speed and accuracy. K-means perform better than k-mean mini-batch because It was giving results faster and easy to implement.

The final phase in this model is associating relations. This phase includes extraction of the mammogram report information through recognizing the named entities then link the extracted relations with the related entities. In order to implement the named entities recognition model for the mammogram reports, it is important to capture the vocabulary of these domain-specific entities. SpaCy library supports very fast statistical entities' recognition, by assigning a label to each contiguous tokens' span [23]. Mammography reporting generally adopts the BI-RADS terminology, which the default spaCy NER model isn't able to identify. All spaCy models support adding a new entity type, so we update a pre-trained model with new entity types based on the BI-RADS terminology. For that, we provide many examples to meaningfully improve the system through manual mapping. For examples, the terms – (architectural distortion, lymph node, mass, microcalcification, calcification) are annotated as finding, the terms – (right, left, bilateral, unilateral) as laterality, the terms – (spiculated, circumscribed, retroareolar) as margins, the terms – (oval, round, irregular) as shape, the terms – (high, low, middle) as density. The NER model is then training on these examples and make a prediction for it. Through this step, the model receives feedback on its prediction in the form of a loss function error gradient that calculates the difference between the training example and the expected output [24]. The greater the differential, the more important the gradient and the model updates. There are different types of loss functions, the NER model use Mean Square Error (MSE) loss function. As the name indicates, the MSE is calculated as the average squared difference between the predictions and the real observations [25]. The following mathematical formulation represent the idea of MSE:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

MSE is concerned only with the average magnitude of the error, regardless of its direction. Nevertheless, as a result of squaring, forecasts far away from real values are heavily penalized relative to less deviated forecasts [25]. Additionally, MSE has good mathematical properties that make it easier to measure gradients. Figure 3shows the NER model process.

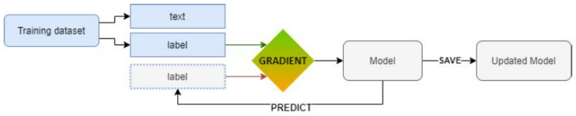
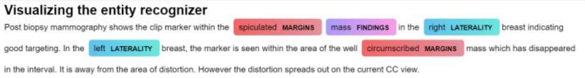


Figure 3: NER model process

Because we don't want the NER model to just memorize the training dataset, the model must process the training dataset in batches and experiment with mini batch sizes and dropout rates - a rate at which individual features and representations are randomly "drop" [26] for a number of training iterations. In order to tune the accuracy for the mammogram NER model, we set the number of iterations=1000 and the dropout rate =0.35 which means that each feature or internal representation has a 0.35 likelihood of being dropped. InFigure 4, we present an output generated by the mammogram NER model on a sample from mammogram reports that contain the labeled named entities with different color base on their class.

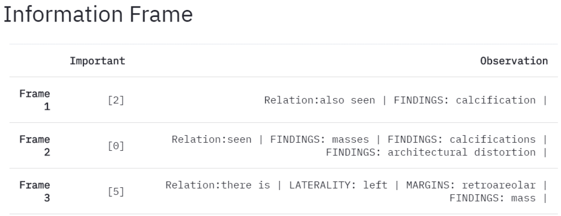


Visualizing the entity recognizer

Post biopsy mammography shows the clip marker within the spiculated MARGINS mass FINDINGS in the right LATERALITY breast indicating good targeting. In the left LATERALITY breast, the marker is seen within the area of the well circumscribed MARGINS mass which has disappeared in the interval. It is away from the area of distortion. However the distortion spreads out on the current CC view.

Figure 4: Visualizing the NER model on an example mammogram report.

For the final step in the last phase of the automatic information extraction procedure, which is linking the extracted relations with the named entities in order to generate the information frame, each inputted report processed independently. Firstly, for each sentence in the report, the relation will be extracted by the relation extraction model if it exists. The relation importance will detect depending on its cluster class, the relation with a higher-class number means it is more important. Then, the named entities in that sentence will be recognized. If the named entities don't contain a 'FINDING' which means, it's not an observation so it will be discarded. After that, the linker will link the relation with the named entity in one frame and processes the next sentence untie finishing the whole report. Finally, the linker will connect all frames in one table and display it as shown inFigure 5.



Information Frame

	Important	Observation
Frame 1	[2]	Relation:also seen | FINDINGS: calcification |
Frame 2	[0]	Relation:seen | FINDINGS: masses | FINDINGS: calcifications | FINDINGS: architectural distortion |
Frame 3	[5]	Relation:there is | LATERALITY: left | MARGINS: retroareolar | FINDINGS: mass |

Figure 5: The Automated Information Frames on an example mammogram report.

IV. SYSTEM ARCHITECTURE AND DESIGN

The system architecture is depicted in Figure 6and is designed to ask the doctor to insert the mammogram report text in a specific box. The entered text transforms into the model. The inputted data goes through several phases to generate the information frame. First, the entered text will go into the preprocessing step, then extract the relations from the report. Second, clustering the relation which extracted. Third, recognize of name entity in the report. Finally, link the relation with the name entity then display the information frame based on relation.

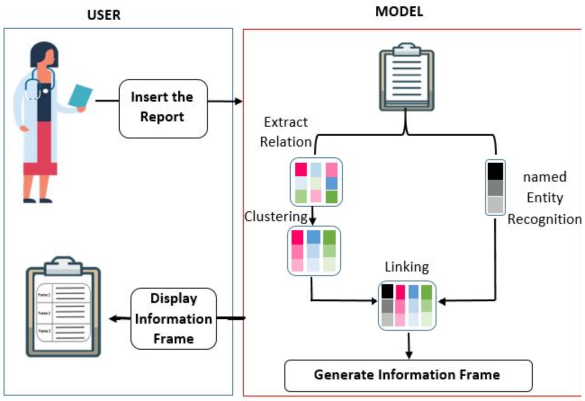


Figure 6: System architecture

The purpose of the system design process is to provide appropriate detailed information about the system and its components so that a consistent and well-running system can be implemented [27]. In order to satisfy that, we use an activity diagram since the visual representations for the system provide a better understanding of the flow of different activities and actions. The activity diagram is not based on what is being generated but on the collection of activities leading to each other and how they are intertwined, with a specific start and finish [28].

The "Automatic information extraction from unstructured mammogram report" has one main screen. It includes two components: writing box to insert the text and button to extract the important information frame from the text. Figure 8 shows the system prototype.

illustrates the automated information extraction system activity diagram.

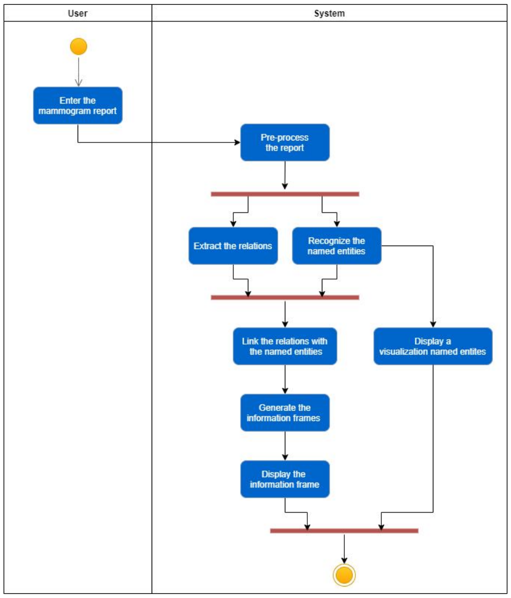
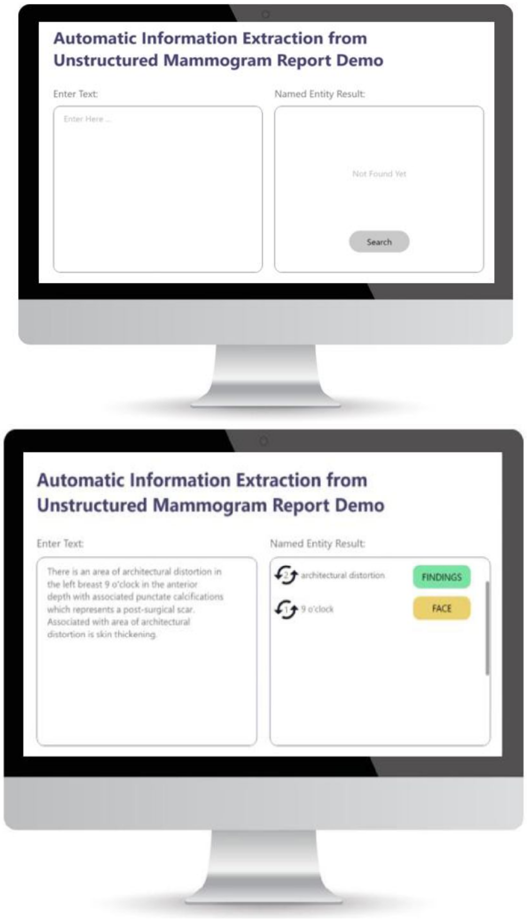

User
System
Enter the mammogram report
Pre-process the report
Extract the relations
Recognize the named entities
Link the relations with the named entities
Display a visualization named entites
Generate the information frames
Display the information frame

Figure 7: The Automated Information Extraction system activity diagram


Automatic Information Extraction from Unstructured Mammogram Report Demo

Enter Text:

Enter Here ...

Named Entity Result:

Not Found Yet

Automatic Information Extraction from Unstructured Mammogram Report Demo

Enter Text:

There is an area of architectural distortion in the left breast 9 o'clock in the anterior depth with associated punctate calcifications which represents a post-surgical scar. Associated with area of architectural distortion is skin thickening.

Named Entity Result:

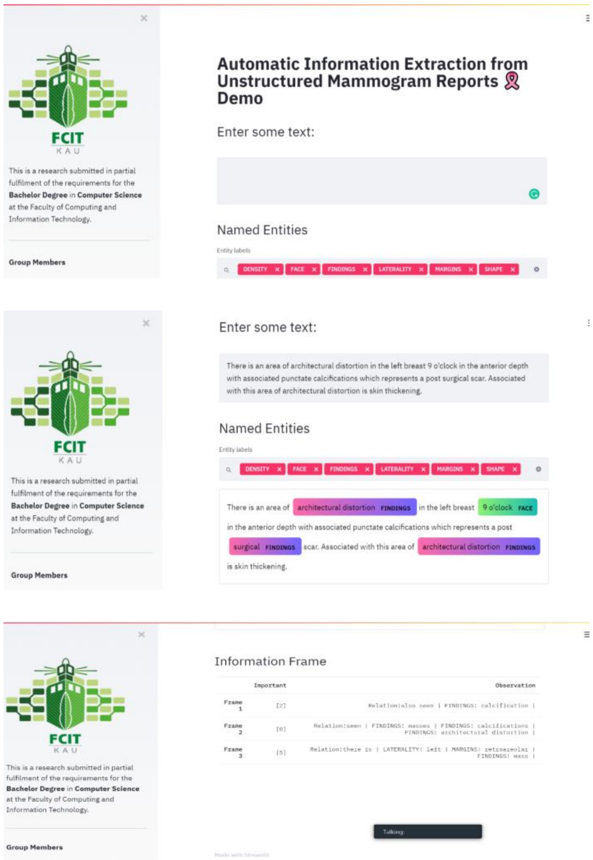
2 architectural distortion FINDINGS

1 9 o'clock FACE

Figure 8: System prototype

Python programming language was used to build the "Automatic Information Extraction" system because of its extensive standard library that provides a lot of useful features. For the purpose of creating an interactive web application for this system, Streamlit framework was used. Streamlit is an open-source app framework for builds web-apps from Python files. It simply converts a python code and a machine learning models with only a couple additional lines of code into a stunning looking application in a quite easy and very interactive way [29]. Streamlit web-app consists of only one component, which will be both the back-end and the front-end. Unlike other web applications that require a back-end where the logic is done and the front-end for the user [29].Streamlit considers as a very powerful framework since it combines a feature of simplicity in writing apps the same way to write plain Python scripts and the feature of rapidly creating and deploying a web-app.

The web-application, as shown in the screenshots of Figure 9, contains initially a box to insert a text ,then after inserting it, the named entity will display and then the information frame is shown.



FCIT
KAU

This is a research submitted in partial fulfilment of the requirements for the **Bachelor Degree** in **Computer Science** at the Faculty of Computing and Information Technology.

Group Members

Automatic Information Extraction from Unstructured Mammogram Reports 🎗 Demo

Enter some text:

Named Entities

Entity labels

DENSITY × FACE × FINDINGS × LATERALITY × MARGINS × SHAPE ×

FCIT
KAU

This is a research submitted in partial fulfilment of the requirements for the **Bachelor Degree** in **Computer Science** at the Faculty of Computing and Information Technology.

Group Members

Enter some text:

There is an area of architectural distortion in the left breast 9 o'clock in the anterior depth with associated punctate calcifications which represents a post surgical scar. Associated with this area of architectural distortion is skin thickening.

Named Entities

Entity labels

DENSITY × FACE × FINDINGS × LATERALITY × MARGINS × SHAPE ×

There is an area of architectural distortion FINDINGS in the left breast 9 o'clock FACE in the anterior depth with associated punctate calcifications which represents a post surgical FINDINGS scar. Associated with this area of architectural distortion FINDINGS is skin thickening.

FCIT
KAU

This is a research submitted in partial fulfilment of the requirements for the **Bachelor Degree** in **Computer Science** at the Faculty of Computing and Information Technology.

Group Members

Information Frame

	Important	Observation
Frame 1	[2]	Relation:also seen | FINDINGS: calcification |
Frame 2	[0]	Relation:seen | FINDINGS: masses | FINDINGS: calcifications | FINDINGS: architectural distortion |
Frame 3	[5]	Relation:there is | LATERALITY: left | MARGINS: retroareolar | FINDINGS: mass |

Talking:

Made with Streamlit

Figure 9: Web application screenshots

V. SYSTEM ASSESSMENT

Quality control was performed to asses the overall analysis of specification, design, and code, as well as to check the software's functionality and accuracy by running it, to detect errors and to estimate its reliability making sure it met the requirements. The laptop hardware specifications used to conduct the experiments were the following: (a) Lenovo with windows 10, Core i7, CPU 2.70 GHz, RAM 8 GB, (b) Dell with windows 10 pro, Core i7, CPU 2.50 GHz, RAM 8 GB, and (c) MacBook Pro, Core i5, CPU 2.30 GHz, RAM 8 GB.

The considered performance measures included a confusion matrix, which is a table often used to define the output of classification modeling on a test dataset when the real values are a priori defined. All the measurements can be determined using the 4 parameters [30]:

(a) True Positive,(TP) representing positive observation and positive prediction.

(b) False Negative,(FN)representing positive observation and negative prediction.

(c) True Negative, (TN) representing negative observation and negative prediction.

(d) False Positive, (FP) representing negative observation and positive prediction.

The four performance measures can be calculated from a confusion matrix and the following metrics can be derived:

(a) Accuracy Metric .

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

(b) Precision Metric

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

(c) Recall Metric

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

(d) F1 score metric

$$F1\ Score = 2 \times \frac{(Recall * Precision)}{(Recall + Precision)}$$

We performed unit testing, to evaluate the accuracy of the system comparing an input-output pair and return a boolean value True or False. In case of True result, the code is consider to behave as expected, whereas in case of a False result the conclusion is the opposite [31]. In order to evaluate the success of the relation extraction model, we used a test set of 30 mammography reports separately to determine the actual relations, single or multiword relation. We applied our relation extraction method within the chosen 30 report samples and compared the metrics of the proposed system against manual manipulations for extracting relation types. We used the confusion matrix for the evaluation of the completeness of the relation extraction model. Table 2 shows the average of confusion matrices in this phase.

TP is the number of correct relations that the model extracted, where the FP is the number of incorrect words that the model extracted as a relation. TN is the number of incorrect relations that the model doesn't extract, on the other hand, the FN is the number of correct relations that the model doesn't extract.

Table 2 : Average of confusion matrices in relation extraction

Relation extraction model		
Total relation = 101	**Actual relation**	
Predict relation	TP=81	FP= 32
	FN= 20	TN= 1867

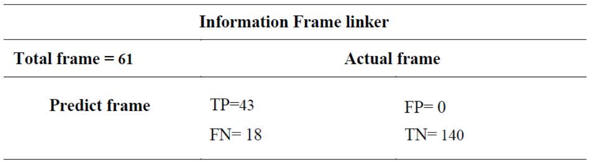
After calculating the performance of each report we calculated the average performance as shown in Table 3.

Table 3: Average performance measures of relation extraction

Accuracy	Precision	Recall	F1 score
93.9 %	62.1 %	83%	68.2 %

An assessment of the success of the final stage in this procedure, we used a test set of 30 mammogram reports in order to test the information frame linker. We applied the information frame linker on the test set and compared the results with the expected information frames. In order to determine the correctness of the information frame linker, we use the confusion matrix. Table 4 shows the performance measurement in this phase.

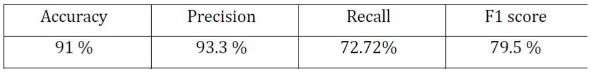
Table 4: Average confusion matrices of information frame linker.



Information Frame linker		
Total frame = 61	**Actual frame**	
Predict frame	TP=43	FP= 0
	FN= 18	TN= 140

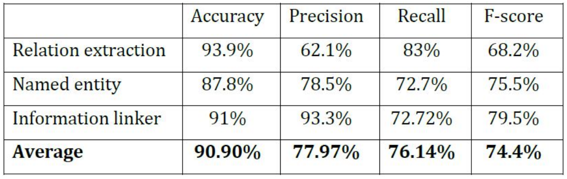
After calculate performance of each report we calculate the average performance as shown in Table 5.

Table 5: Performance measurement for Information Frame linker.



Accuracy	Precision	Recall	F1 score
91 %	93.3 %	72.72%	79.5 %

Table 6 : System test result



	Accuracy	Precision	Recall	F-score
Relation extraction	93.9%	62.1%	83%	68.2%
Named entity	87.8%	78.5%	72.7%	75.5%
Information linker	91%	93.3%	72.72%	79.5%
Average	**90.90%**	**77.97%**	**76.14%**	**74.4%**

As described previously, the unit testing was done for each phase in order to indicate that each model is behaving as intended. But for the whole system, in order to test its performance, we calculate the average of the summation of the performances of each model. Table 6 indicates the system test result.

The 'Automatic Information Extraction from Unstructured Mammogram Reports' system was able to extract some relation either it is single or multi-word relation, recognize a specific named, generate the information frame that linker the relation with the observation. This system thus is advantageous and requires much less effort than manual reviewing for the report. The system performance that presented early was constrained by different limitations. First, the relations extraction model used the Spacy POS tagger to extract the relations, which is not necessarily right when interpreting the mammogram report because of that the narrative form of the mammogram report does not necessarily obey the syntactic rules. Secondly, even though the performance of the named entity recognition model reaches 87.8% for accuracy and 75.5% for the F1 score it could be higher if we were able to obtain more labeled data. However, the performance for the whole system seems to be reasonably captured based on the F1 score of 74.4%. Ultimately we will need a much wider collection of data to prove our conclusions are compelling and generalizable.

VI. CONCLUSIONS

The paper investigated and implemented a system for Automatic Information Extraction from Unstructured Mammogram Reports. The system included a data pre-processing stage that was followed by three 3 phases. The first phase involved relation extraction by using a spacy library. In the second phase, K-means was used by the Scikit-learn library to perform clustering based on the identified relation. The third phase associated the relations by using named entity recognition to generate the information frame. The system was trained and tested by using Python Language with spacy and Scikit-learn libraries. The system obtains an accuracy of 90.9%, Precision 77.97 %, Recall 76.14 % and F1-score 74.4%.

REFERENCES

[1] F.Bray, J.Ferlay, I.Soerjomataram, R.L.Siegel, L.A.Torre, A.Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, 68(6), 394-424. doi:10.3322/caac.21492, 2018

[2] A.E.Ahmed, D.K.McClish, T.Alghamdi, A.Alshehri, Y.Aljahdali, K.Aburayah, A.R.Jazieh, “Modeling risk assessment for breast cancer in symptomatic women: a Saudi Arabian study,” Cancer management and research, 11, 1125-1132. doi:10.2147/CMAR.S189883, 2019.

[3] F.AlMulhim, A.Syed, W.Bagatadah, A.AlMuhanna, “Breast cancer screening programme: experience from Eastern province, Saudi Arabia,” Eastern Mediterranean Health Journal, 21(2), 2015.

[4] H.Greenlee, M.J.DuPont□Reyes, L.G.Balneaves, L.E.Carlson, M.R.Cohen, G.Deng, S.M.Zick, “Clinical practice guidelines on the evidence□based use of integrative therapies during and after breast cancer treatment,” CA: A Cancer Journal for Clinicians, 67(3), 194-232, 2017.

[5] R.M.Alotaibi, H.R.Rezk, C.I.Juliana, C.Guure, “Breast cancer mortality in Saudi Arabia: Modelling observed and unobserved factors,” PloS one, 13(10), e0206148, 2018.

[6] S. A. Ministry of Health, “Early detection of breast cancer,” [Online https://www.moh.gov.sa/Ministry/Projects/breast-cancer/Pages/jeddah.aspx, 2019, May 3]

[7] M.Beslaim, I.Baroum, B.Salman, B.Baghlaf, M.Al-Farsi, “Breast Cancer Screening Program in Jeddah, secSaudi Arabia: Is There a Need for a National Program?,” Int J Womens Health Wellness, 2, 042, 2016

[8] A.M.Sheikh “Knowledge of family medicine residents regarding the interpretation of chest X-rays in Jeddah, Saudi Arabia,” International Journal of Medical Science and Public Health, 7(7), 538-545, 2018.

[9] N.Thiebaut, A.Simoulin, K.Neuberger, I.Ibnouhsein, N.Bousquet, N.Reix, C.Mathelin, “An innovative solution for breast cancer textual big data analysis,” arXiv preprint arXiv:1712.022592017, 2017.

[10] R.Weegar, H.Dalianis, “Creating a rule based system for text mining of Norwegian breast cancer pathology reports,” Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, 2015.

[11] W.Sun, Z.Cai, Y.Li, F.Liu, S.Fang, G.Wang, “Data processing and text mining technologies on electronic medical records: a review,” Journal of healthcare engineering, 2018.

[12] H.Gao, E.J.A.Bowles, D.Carrell, D.S.Buist, “Using natural language processing to extract mammographic findings,” . Journal of biomedical informatics, 54, 77-84, 2015.

[13] A.Yala, R.Barzilay, L.Salama, M.Griffin, G.Sollender, A.Bardia, F.Polubriaginof, "Using machine learning to parse breast pathology reports," Breast cancer research and treatment, 161(2), 203-211, 2017.

[14] S.M.Castro, E.Tseytlin, O.Medvedeva, K.Mitchell, S.Visweswaran, T.Bekhuis, R.S.Jacobson, "Automated annotation and classification of BI-RADS assessment from radiology reports", Journal of biomedical informatics, 69, 177-187, 2017.

[15] C.Moore, A.Farrag, E.Ashkin, "Using natural language processing to extract abnormal results from cancer screening reports," Journal of patient safety, 13(3), 138, 2017.

[16] A.Gupta, I.Banerjee, D.L.Rubin, "Automatic information extraction from unstructured mammography reports using distributed semantics," Journal of biomedical informatics, 78, 78-86, 2018.

[17] G.Mahadevaiah, A.Hiremath, V.Agarwal, P.Kumaraguru, A.Dekker, "Automating data mining of medical reports," International Journal of Computer Science and Technology (IJCST), Vol 1, 2019.

[18] I.Banerjee, S.Bozkurt, E.Alkim, H.Sagreiya, A.W.Kurian, D.L.Rubin, "Automatic inference of BI-RADS final assessment categories from narrative mammography report findings," Journal of biomedical informatics, 92, 103137, 2019.

[19] I.Guyon, S.Gunn, M.Nikravesh, L.A. Zadeh, "Feature extraction: foundations and applications," Vol. 207: Springer. (2008).

[20] F.Pedregosa, G.Varoquaux, A.Gramfort, V.Michel, B.Thirion, O.Grisel, , V. Dubourg, "Scikit-learn: Machine learning in Python," Journal of machine learning research, 12(Oct), 2825-2830, 2011.

[21] J.Béjar Alonso, "K-means vs Mini Batch K-means: A comparison," 2013.

[22] K.Sirait, E.B.Nababan, "K-Means algorithm performance analysis with determining the value of starting centroid with random and KD-tree method," Journal of Physics: Conference Series. 2017.

[23] R.Jiang, R.E.Banchs, H.Li, "Evaluating and combining name entity recognition systems," Proceedings of the Sixth Named Entity Workshop, 2016.

[24] K.P.Körding, D.M.Wolpert, "The loss function of sensorimotor learning," Proceedings of the National Academy of Sciences, 101(26), 9839-9842, 2004.

[25] R.Reed, R.J.Marks II, "Neural smithing: supervised learning in feedforward artificial neural networks," MIT Press, 1999.

[26] H.Xu, S.Li, R.Hu, S.Li, S.Gao, "From random to supervised: A novel dropout mechanism integrated with global information," arXiv preprint arXiv:1808.08149, 2018.

[27] A.Kossiakoff, W.N.Sweet, S.J.Seymour, S.M.Biemer, "Systems engineering principles and practice," (Vol. 83): John Wiley & Sons, 2011

[28] R.M.Bastos, D.D.A.Ruiz, "Extending UML activity diagram for workflow modeling in production systems," Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002.

[29] T.Adrien, A.Treuille, "Streamlit," [Online https://www. streamlit.io/, 2020]

[30] E.Alpaydin, "Introduction to machine learning", MIT press, 2020.

[31] G.P.Sarma, T.W.Jacobs, M.D.Watts, S.V.Ghayoomie, S.D.Larson, R. C.Gerkin, "Unit testing, model validation, and biological simulation,", F1000 Research, 5. 2016.