The University of Tehran Press

Interdisciplinary Journal of Management Studies (IJMS)
Home Page: https://ijms.ut.ac.ir

Online ISSN: 2981-0795

Cancer detection from textual data using a combination of machine learning approach

Bita Salmanpoursohi[1] | Amir Daneshvar[2*] | Shakiba Salmanpoursohi[3] | Adel Pourghader Chobar[4] | Fariba Salahi[5]

1. Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: bita.salmanpour@gmail.com
2. Corresponding Author, Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: a_daneshvar@iauec.ac.ir
3. Department of Information Technology Management, Tehran North Branch, Islamic Azad University, Tehran, Iran. Email: sh.salmanpour@gmail.com
4. Department of Industrial Engineering, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran. Email: apourghader@gmail.com
5. Department of Industrial Management, Tehran South Branch, Islamic Azad University, Tehran, Iran. Email: f_salahi@iauec.ac.ir

ARTICLE INFO

Article type:
Research Article

Article History:
Received 14 July 2023
Revised 14 November 2023
Accepted 18 November 2023
Published Online 12 June 2024

Keywords:
Logistic Regression,
Naive Bayes,
Random Forest,
Support Vector Machine,
Cancer Detection.

ABSTRACT

Recently, cancer has become one of the main diseases and causes of death of people all over the world. For this purpose, extensive research has been done on the prediction and early detection of this disease in the body of patients in different fields. Artificial intelligence and data mining approaches are among the methods that have helped researchers in diagnosing this disease. In this research, a machine learning approach for early and timely diagnosis of cancer disease is presented. For this purpose, it uses logistic regression techniques, Naive Bayes, two versions of Random Forest and Support Vector Machine, which work in parallel with each other. As a result of the integration of the techniques, the proposed system achieves higher accuracy and reduces errors compared to the basic methods. The performance of the proposed method was evaluated using different criteria and showed superior results compared to traditional methods.

Cite this article: Salmanpoursohi, B.; Daneshvar, A.; Salmanpoursohi, Sh.; Pourghader Chobar, A. & Salahi, F. (2024). Cancer detection from textual data using a combination of machine learning approaches. *Interdisciplinary Journal of Management Studies* (*IJMS*), 17 (3), 1001-1014. DOI: http//doi.org/10.22059/ijms.2023.362252.676037

Publisher: University of Tehran Press.
DOI: http//doi.org/10.22059/ijms.2023.362252.676037

1. Introduction

Cancer is a global public health concern, emerging as a leading cause of death (Torre et al., 2015). Uncontrolled cell growth characterizes this disease, substantially impacting human survival rates. Early detection of cancer remains a pivotal challenge in the medical field (Hu et al., 2018). Over the past two decades, increased throughput of data from various medical sources has made cancer a data-driven science (Lisboa et al., 2010). However, in many cases, recording reports in different types of medical records in the form of free texts is still a norm (Harkema et al., 2011). Even if reporting systems in the future are more efficient at collecting structured data, the problem of managing old data in free text format will not be solved because past information sources are very valuable.

One of the best solutions available to solve this problem is the automatic extraction of this information. Currently, text mining systems that operate automatically perform processing on texts and data that are created by humans and do not have a specific structure. In this process, natural language processing methods are used, and finally, the output is knowledge extraction from the available data. (Hosseini et al. 2022; Conceição & Couto, 2021). Natural language processing techniques, knowledge management methods, artificial intelligenceet al., and machine learning, are used for the efficient processing of large sets of documents, information retrieval, document classification (appropriate categorization based on content), information extraction, etc. (Soffer et al., 2020; Chobar et al. 2022). Using of artificial intelligence approach like artificial intelligence approaches such as machine learning (ML) and data mining (DM) can be used in different parts of the field of treatment and their management. Among these things, we can mention searching for the required information, collecting and managing patients' information, how to diagnose and predict diseases, and checking the feedback of patients' treatment results. The use of these items can ultimately improve the performance of specialists and the effective treatment of patients. In addition, extracting the knowledge available in a large amount of recorded data can reduce a significant part of the costs of the treatment system, while in many cases, it will lead to accelerating the treatment process and increasing the accuracy of diagnosis and treatment of diseases.

From the management point of view, it is very important to have a possibility to diagnose cancer from textual data. With access to a large amount of textual data collected about diseases, health system managers can propose an artificial intelligence-based mechanism for cancer diagnosis. This system can also be used as an assistant to a specialist to help a specialist physician in the field of diagnosis. Managers can also consider new methods of teaching medical students by using artificial intelligence-based solutions for disease detection. By using methods based on artificial intelligence, the education process can be carried out at a lower cost. The main question and purpose of this research are to find a way to achieve good results for cancer detection by using text data. To get the answer to this question, we use machine learning methods and textual data preprocessing. In the presented model of this research, first, the data required for cancer diagnosis are extracted from the text format using the text mining technique and are placed in the structured data format. Next, a cancer diagnosis is performed by applying machine learning methods on a cancer diagnosis is performed by applying machine learning methods to the extracted data. In the continuation of this research, the use of ensemble techniques is suggested to improve the overall accuracy accuracy of prediction. In ensemble techniques, more than one model is used to the data to reduce the overall error rate. In fact, this issue is synonymous with something we applyet al. et al. constantly apply in our daily lives, such as asking for the suggestions of several experts before making an important decision to reduce the possibility of a wrong decision (Ahmad et al., 2020; Asgharizadeh et al. 2022). There are different aggregation methods for solving problems, each of which works differently. One learning algorithm is used for methods such as bagging and boosting. In boosting-based methods, data sets are randomly generated, but in bagging-based methods, data are weighted and do not have the same probability of selection. Methods based on stacking and voting by combining several classification algorithms usually show better performance (Upadhyay et al.et al., 2021; Jahangiri et al., 2021). In the current study, the method based on majority voting is applied to solve the problem of cancer detection. In this research, in order to achieve the desired results faster, the desired methods will be implemented in a distributed manner in Spark format.

2. Literature review

Iqbal et al. (2021), prostate image features were extracted using 10-fold cross-validation. Subsequently, a classification method based on ResNet and Deep Learning LSTM achieved an impressive accuracy of 99.84%. Alternatively, Erdem & Bozkurt (2021) examined the performance of different ML approaches, such as KNN, SVM, RF, logistic regression, LR, NB, linear discrimination analysis, linear classification, MLP, and deep neural network, for prostate cancer forecast. The researchers utilized available online patient data to implement these methods. Remarkably, MLP recorded the highest performance with an accuracy of 97%. In the research of Liew and his colleagues (Liew et al. Liew and his colleagues (Liew et al., 2021), a new technique was developed for breast cancer categorization that uses a combination of deep learning and XGBoost. This technique was applied to the BreaKHis data set and achieved a good accuracy of 97% in the classification of breast cancer images. In Mahesh et al.'s (2022) study, an ML method was proposed to predict breast cancer. The researchers addressed the issue of imbalanced data by employing the SMOTE technique. Subsequently, Naive Bayes, decision trees, Random Forest, and cumulative methods were utilized for data classification. The XGBoost and Random Forest methods, based on cumulative techniques, achieved the highest accuracy of 98.20%, according to the experimental results. Recently, many ML methodset al. et al.et al. et al. methods have been suggested for the early detection of various types of cancer, including breast cancer (Aldhaeebi et al., 2020), skin cancer (Dildar et al., 2021), lung cancer (Riquelme & Akhloufi, 2020), etc. In the following, a number of recent works that were presented for cancer detection in different fields are reviewed. For example, Bhatia et al. (2019) presented an approach for lung cancer discovery using ML. For this purpose, deep learning was employed to extract features from the approach. Moreover, a combination of multiple classifiers was utilized for cancer prediction, with the XGBoost and Random Forest ensemble showing the best performance, obtaining an accuracy of 84%. Nanglia et al. (2021), a hybrid Feed-Forward Back Propagation neural network (FFBPN), was employed for lung cancer detection. Within this context, a fusion of SVM and FFBPN was implemented to create a hybrid mechanism, reducing the computational complexity associated with classification. The suggested research obtained a system accuracy of 98.08%. Zhu et al. (2016) utilized deep convolutional neural networks (DCNNs) and direct examination of pathological images to forecast the survival time of lung cancer disease. Hua et al. (2015) employed an ML approach to classify pulmonary nodules in 2D CT images, training two deep end-to-end models, namely DBN and CNN, on raw lung images. In the article by Khorshid et al.et al. (2021), a comparative analysis of five different classification algorithms, including LR, SVM, K-NN, Weighted KNN, and Gaussian NB, was performed on the breast cancer dataset. The dataset is taken from the UCI website. The primary objective of this study is to use the ML approach to the classification of breast cancer in women. The findings demonstrated that among all classifiers, K-NN displayed the highest accuracy, reaching 96.7%. In the research conducted by Mojrian et al. (2020), a multi-layer fuzzy expert system based on an extreme learning machine (ELM) with a radial basis function (RBF), referred to as ELM-RBF, was proposed for breast cancer detection. A linear SVM model was also employed for comparison, yielding satisfactory results. In another study by Mushtaq et al. (2020), a solution for breast cancer detection was presented, where the performance of the KNN model was assessed using different distance functions and varying values for K. The experiments involved three iterations with distinct selection mechanisms. The findings revealed that the KNN model attained the highest accuracy when utilizing the chi-squared-based feature selection in conjunction with the Manhattan distance function. In the study conducted by Kaur et al. (2021), a deep learning feature engineering model incorporating an optimized Xg-boost classifier was employed for skin cell image classification and skin cancer detection. The Xg-boost structure was optimized using gray wolf optimization. The results showcased accuracy and precision values of 98.34% and 97.35%, respectively. Garg and colleagues (2021) aimed to propose a system that integrates image processing and a deep learning model, specifically a convolutional neural network (CNN), for skin cancer detection. They utilized the Transfer Learning method to enhance image classification accuracy, resulting in an accuracy rate of 90.51% for their proposed CNN model. In the research paper by Hekler et al. (2019), a methodology was presented that combined the capabilities of convolutional neural networks (CNN) with human knowledge to classify suspicious skin cancer images. This integration achieved an accuracy of 82.95%, surpassing the accuracy achieved by AI or humans alone, which were 81.59% and 42.94%, respectively. In the article

(Botlagunta ., 2023), a system was developed that detects breast cancer. In this system, the information in the patient's medical records is extracted using text mining and data processing techniques. Welch Unpaired t-test was applied to calculate the meaning of the data, and finally, ML methods were used to categorize the data, and the decision tree was able to reach 83% accuracy. In the paper (Hjaltelin et al., 2023), researchers used information from symptom codes in a Danish patient database recorded over 42 years to diagnose pancreatic cancer symptoms. In addition, for a more comprehensive comparison and the possibility of making a comparison, the information available in the electronic health records was also used. The data mining of these two intelligence sources simultaneously led to the detection and registration of early symptoms of pancreatic cancer.

In this section, researches based on ML approach for the diagnosis of lung, breast, prostate and skin cancers were reviewed. At the end of this section, a summary of related works and the results obtained from them are given in Table 1.

Table 1. categorization of the related work

Research	Field	Method	Achievement
Nanglia et al, 2021	Lung cancer	hybrid Feed-Forward Back Propagation neural network	Accuracy 98.08%
Zhu et al., 2016	Lung cancer	Deep convolutional neural networks	Concordance index 0.629
Hua et al., 2015	Lung cancer	DBN	Specificity 82.2%
Bhatia et al, 2019	Lung cancer	XGBoost and Random Forest	Accuracy 84%.
Khorshid et al, 2021	Breast cancer	K-NN	Accuracy 96.7
Mojrian et al., 2020	Breast cancer	ELM- RBF	R^2 0.9374
Mushtaq et al, 2020	Breast cancer	KNN	Accuracy 99.42%
Liew et al, 2021	Breast cancer	deep learning and XGBoost	Accuracy 97%
Mahesh et al, 2022	Breast cancer	XGBoost and Random Forest + SMOTE	Accuracy 98.20
Iqbal et al, 2021	Prostate cancer	ResNet & Deep Learning LSTM	Accuracy 99.84%
Erdem & Bozkurt, 2021	Prostate cancer	MLP	97% accuracy
Kaur & Kaur, 2021	Skin cancer	Xgboost + gray wolf optimization	Accuracy 98.34%
Garg et al, 2021	Skin cancer	Deep learning + convolutional neural network	Accuracy 90.51%
Hekler et al., 2019	Skin cancer	CNN	Accuracy 82.95%
Botlagunta et al, 2023	Breast cancer	Text mining with Decision tree	Accuracy 83%
Hjaltelin et al, 2023	Pancreatic cancer	Statistical analysis of Text mining results	haemorrhages (p-value = $4.80 \cdot 10^{-08}$) and headache (p-value = $2.12 \cdot 10^{-06}$)

3. The proposed system

Any ML algorithm is able to be used as a data set and trained and used to predict new data. The efficiency of the ML approach, in addition to its parameter settings, depends on the data set of the problem. Therefore, it is not easy to say which method can be suitable for solving the problem. By using ensemble techniques, several different methods can be trained on the problem data set and solve existing problems.

Current research utilizes various methods for building a cancer detection system, consist of logistic regression, Naive Bayes, two versions of random forest, support vector machine, and text mining techniques. The proposed system combines these modeling methods using the majority voting method. The subsequent section provides detailed explanations of each mentioned approach, while Figure 1 shows the flowchart of the suggested system.

Since the correct setting of the parameters of each learning model has an effective role in the efficiency of that model and the performance of the overall system, this operation has become one of the main parts of the system, as shown in the flowchart of the proposed method. A successive halving method was used to do this. The search process in the first iteration evaluates all candidates (N different combinations of parameters) with a small budget (e.g. number of training samples). If B is the total budget allocated to all candidates, each candidate receives a B/N budget. At the end of each

iteration, half of the best candidates are selected for the next iteration, and more resources are allocated to them. Only a subset of candidates survives the next iterations, and finally, one candidate is selected.

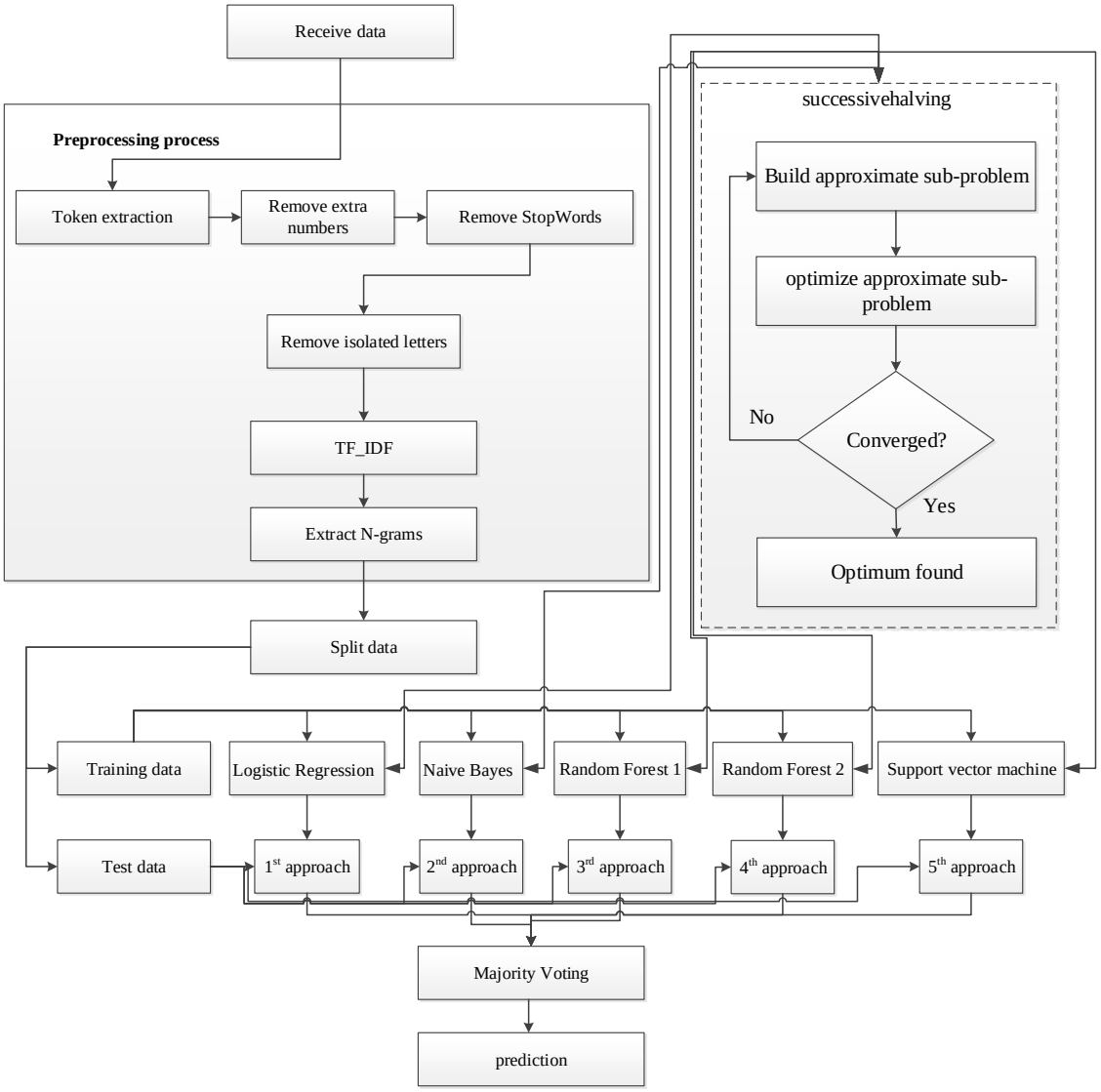

Fig. 1. Flowchart of the proposed system

3.1. Data Preprocessing

Fields like medicine, which rely on accurate information for predictions, can leverage advancements in text mining methods to extract relationships. For this purpose, this study aims to utilize text mining techniques in order to identify relationships within relevant data and extract valuable insights. In this way, the information presented in the form of text will be transformed into structured data in this research.

The initial stage of the proposed approach involves text preprocessing. During this step, the existing textual data is tokenized and stop words are eliminated. Additionally, words that frequently appear in any document and hold no significant meaning are discarded. Following these operations, the Case Folding operation is carried out. Here, all words are examined for their lowercase or uppercase form, and if a word appears multiple times in different cases, it is treated as a single entity during the modeling process (Agarwal and Jhai, 2012).

In the subsequent stage, stemming is employed to normalize word forms within the documents. Rooting techniques are utilized to analyze different words based on their meanings. Words that possess similar meanings but differ in appearance are grouped together and considered as a feature

(Ramasubramanian and Ramya, 2013). Eventually, IDF-TF weighting methods are applied to determine the appropriate weight for each word. In this technique, each word is assigned a weight proportional to its frequency within each document and across all documents. This weight is calculated using Equation 1, where t_k denotes the $k-th$word word, di represent the $i-th$ document, NDenotes the total number of documents and d_kIgnifies the number of documents containing the term t_k (Weiss et al., 2010).

$$TFIDF\left(t_k, d_i\right) = TF\left(t_k, d_i\right) \times \log\left(\frac{N}{d_k}\right) \quad (1)$$

3.2. Modeling

In this section, the details of the used algorithms are explained. Some of the algorithms used are Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine.

3.2.1. Logistic Regression

Logistic regression is a statistical model that determines the correlation between the dependent variable and independent variables using available data and observations (Saadi & Abolfazl, 2000). The logistic regression technique is a multivariate analysis that considers all the predictors of an issue simultaneously. The logistic regression model is a unique type of regression model in which the dependent variable is two states and takes only values of zero or one.

3.2.2. Naive Bayes (NB)

Using the NB classifier, we assume that the values of the objective functions are independent of each other. That is, the probability of observing the link of attributes$a_1, a_2 \ldots a_n$, according to the target value of the sample, is equal to the equations (2-3)

$$P(a_1, a_2, \ldots., a_n|v_j) = argmax_{v_j \in V} \prod P\left(a_i|v_j\right) \quad (2)$$

$$v_{NB} = argmax_{v_j \in V}\ P\left(v_j\right) \prod P\left(a_i|v_j\right) \quad (3)$$

The Naive Bayes classifier simplifies the problem by assuming that the inputs are independent of each other, which reduces a multivariate problem to a set of univariate problems. In this approach, the number of distinct terms $P(a_i|v_j)$That needs to be approximated from the training data $P(a_1, a_2, \ldots., a_n|v_j)$is much smaller than the estimated terms required by Bayesian theory (Islam et al., 2007).

3.2.3. Random Forest

Classification can be achieved with decent accuracy using any decision tree algorithm. Random forests, in other words, are made up of multiple decision trees, which collectively make more accurate decisions. When constructing a decision tree, even a slight modification in the learning patterns can lead to significant changes in its structure. To address this issue, random forests come into play by averaging the results of all decision trees. Notably, random forests excel at assessing the importance of variables and determining their predictive role (Braiman, 2001).

3.2.4. Support Vector Machine (SVM)

SVM is an approach used to classify patterns, whether they are linear or non-linear. It falls under supervised learning and is capable of addressing both regression and classification problems, particularly those involving binary classification, with satisfactory accuracy. The algorithm works by creating an optimal hyperplanehyperplanehyperplane, which is considered optimized when it maximizes the distance between the hyperplane and the nearest point for each pattern. A notable advantage of SVM is its effective performance in large spaces (Pradhan, 2012).

3.2.5. Majority voting

The classifier ensemble based on majority voting is a valuable approach to minimize minimize obtaining multi-expert recommendations and minimizing the risk of misdiagnosis (Chandra et al., 2021; Jahangiri et al., 2023). Various learning methods exist, each with its own strengths in specific

domains. However, there is no universally superior algorithm, and the selection of models depends on the data and problem at hand. Rather than selecting a single algorithm, combining the outputs of multiple classifiers is preferred. By leveraging the diverse strengths and weaknesses of these classifiers, a collaborative approach can lead to improved outcomes. Voting algorithms are commonly employed to achieve masking error-masking capabilities and enhance accuracy and reliability in numerous practical and research settings. The concept behind majority voting is that collective judgment outperforms individual judgment. The subsequent section presents the pseudo-code for the cumulative majority voting algorithm.

```
Pseudo code: Majority vote ensemble algorithm
Dataset Train=(Z,K),
Dataset Test=(z,k)
m = size of Test Dataset
n = The number of Classifiers C // 4 in this example
begin
        For i =1, …, n do
                p_i =train classifier (c_i)
        end
        for i=1, …, m do
                for j=1, …, n do
                        Use classifier (cj) to sample x_i
                Finish
                Y_i= maximum votes
        Finish
Finish
```

4. Dataset

The dataset used in this research is taken from a research article (Ye et al., 2016). In this article, to show the effectiveness of the presented solution, the information related to breast, prostate, and lung cancer diseases was extracted from thousands of articles and used. Other researchers have made this dataset available for use.

5. Results

In the following tables, the details of some parameters of the used methods are stated (sci-kit-learn). As previously stated, in the proposed system of this research, the optimal values for the parameters of the basic methods are done using the successive halving method.

In table (3), details of Naive Bayes parameter is shown. The only parameter that we work on it, is "var_smothing", this parameter represents the fraction of the maximum variance across all features, which serves to enhance stability in calculation.

Table 2. Details of Logistic Regression parameters

Parameter	Description	Value
penalty	Specify the norm of the penalty	L2
top	Stopping criteria tolerance	1e-4
C	Regularization strength (positive float) and it's inverse; higher values imply weaker regularization similar to support vector machines	1.0
solver	Choice of algorithm for the optimization problem	blogs
max_iter	The maximum number of iterations required for the solvers to converge	100

Table 3. Details of Naive Bayes

Parameter	Description	value
var_smoothing	The fraction of the maximum variance among all features contributes to the stability of variance calculations.	1e-9

In table (4) and table (5), details of Random Forest parameters are shown. Different parameters like "n_estimators," "min_samples_split," etc. "n_estimators," "min_samples_split," etc. are set for the random forest. Values and descriptions for these parameters are shown in the tables.

Table 4. details of Random Forest 1

Parameter	Description	value
n_estimators	The count of trees present in the forest	100
criterion	The method for evaluating the split's effectiveness	gini
min_samples_split	The minimum sample count needed for internal node splitting.	2
min_samples_leaf	The lowest sample count needed for a leaf node	1

Table 5. details of Random Forest 2

Parameter	Description	value
n_estimators	Number of trees in the forest	150
criterion	Quality measurement function for a split	gini
min_samples_split	The lowest required samples to split an internal node	3
min_samples_leaf	The lowest sample count needed for a leaf node	1

In table (6), details of SVM parameters are shown. The "C" parameter is Regularization parameter that set to 1.0. Different type of kernel can be used in SVM model. We use radial basis function kernel or RBF kernel.

The confusion matrix achieved from the proposed system in train data for cancer detection is shown in Figure 2. According to the results obtained from this matrix, the number of wrongly diagnosed cases in all three classes is not very high and is acceptable. This problem depict the proper performance of the system suggested in this research in all three classes.

Table 6. details of SVM

Parameter	Description	Value
C	Inverse relationship between regularization strength and C.	1.0
kernel	Determines the type of kernel to be used.	Rbf
tol	Specifies the tolerance for the stopping criterion	1e-3

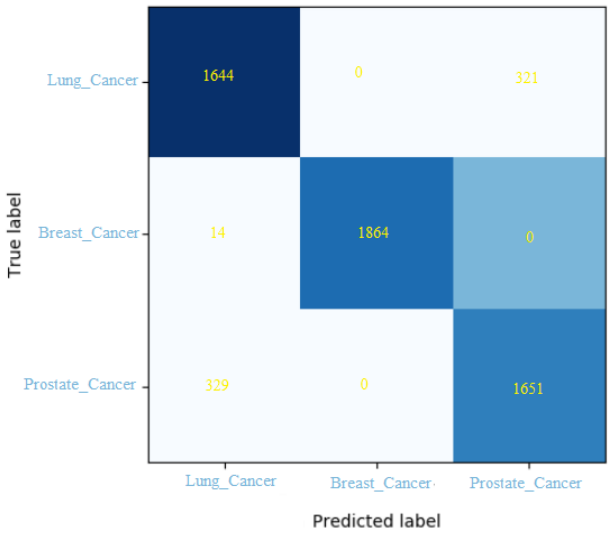

Fig. 2. Confusion matrix obtained from the proposed system

The confusion matrix shows the distribution of correct and incorrect predictions in a dataset. The actual values are displayed in the rows, while the predicted values are represented in the columns. Each data sample can fall into one of four possible states.

- True Positive (TP) denotes the accurate classification of a sample as a member of a specific class.
- False Negative (FN) corresponds to misclassifying a sample from a specific class as belonging to other classes.

- True Negative (TN) represents correct recognition of a sample not belonging to a specific class.
- False Positive (FP) occurs when a sample that does not belong to a specific class is incorrectly labeled as a member of that class.

In this case, the evaluation criteria for each class are calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F - measure = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \tag{7}$$

For each class, TP, FN, FP and TN values will be obtained according to the figure 3. According to the values in the confusion matrix for each class, the evaluation criteria for the training data are calculated. The precision, recall, F-value, and accuracy values for the lung cancer, breast cancer, and prostate cancer classes were recorded in table 7.

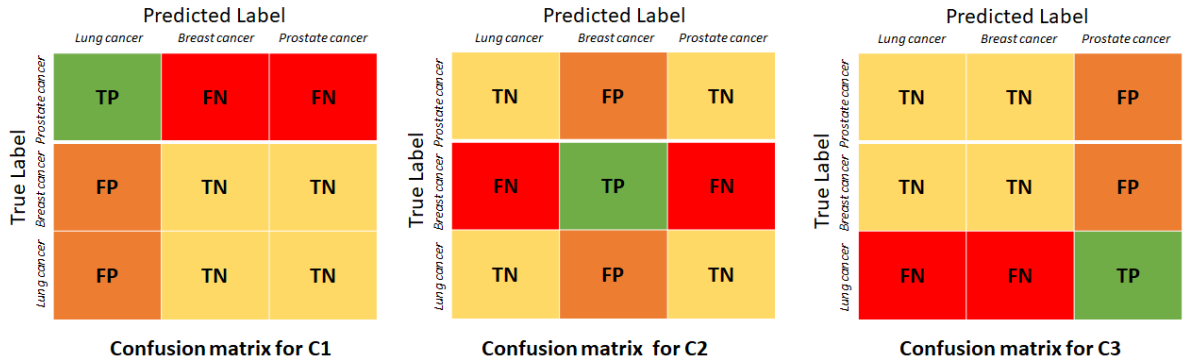

Fig. 3. Confusion matrix for each class

Table 7. precision, recall, F-value, and accuracy for lung, breast, and prostate cancer class

	Lung cancer	**Breast cancer**	**Prostate cancer**
Precision	82.73	100.0	83.76
Recall	83.66	99.25	83.38
F-measure	83.19	99.62	83.56
Accuracy	88.59	99.75	88.83

Due to the fact that in this researchet al., the data collection presented in the research of Ye and his colleagues (Ye et al., 2016) was used, this article was also chosen to compare the findings. The dataset utilized in this research has been employed in numerous studies, including Patel et al. (2022). These studies compare the performance of the suggested method with basic techniques like SVM, Naïve Bayes, and logistic regression without comparing it to other studies. Considering that the method of sampling and dividing the data into training and test data various from the previous methods, the results recorded in these researches cannot be used directly for comparison with the suggested method. The sampling and division of the data will have a great impact on the system's final result. Therefore, in this research, the results of the mentioned basic methods and random forest method were re-examined with the data used in this research, according to the conditions of the implementation of this research. The recorded results for the basic methods obtained using the sampling method of this research have been compared with the findings of the suggested method.

In the following, the findings of the suggested model implementation are compared with SVM, NB, logistic regression, and random forest. Figures (4), (5), (6), and (7) present the results for different criteria obtained from both the proposed system and the basic classifications for better comparison.

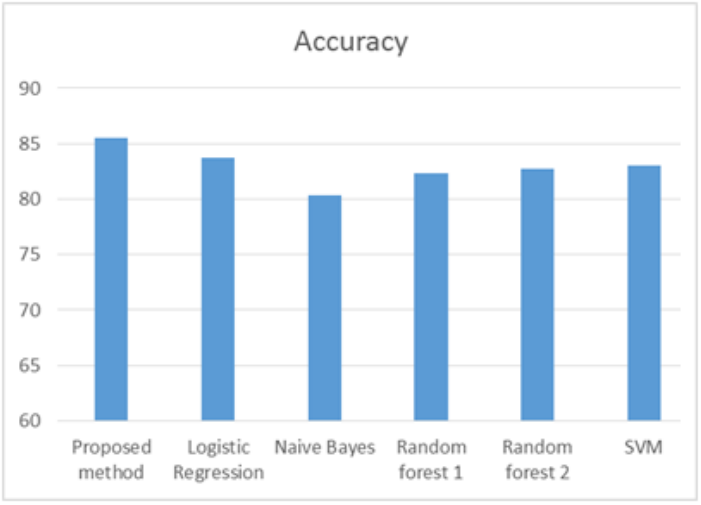

Fig. 4. Comparison of various models according to accuracy criteria

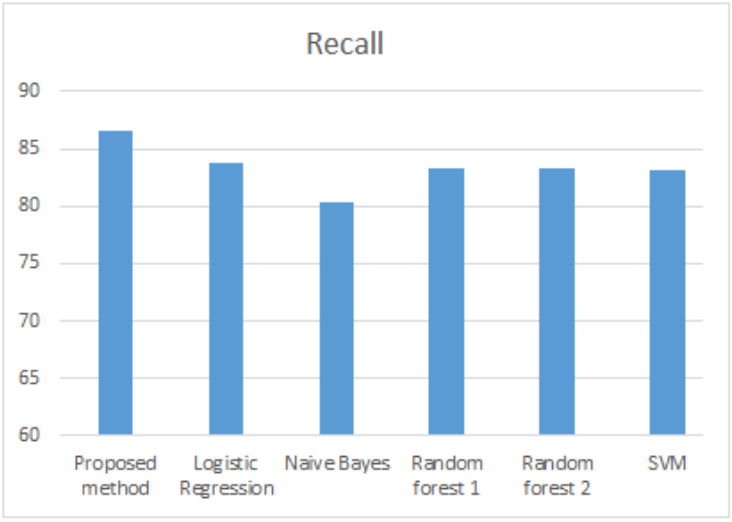

Fig. 5. Comparison of various models according to recall criteria

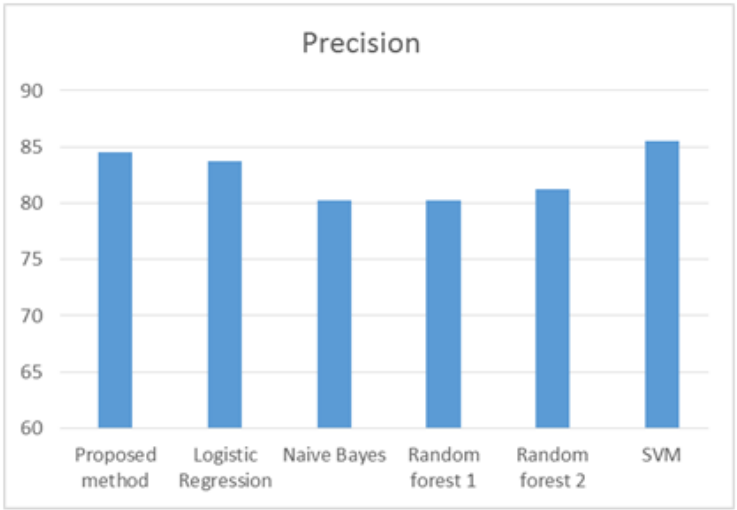

Fig. 6. Comparison of various models according to precision criteria

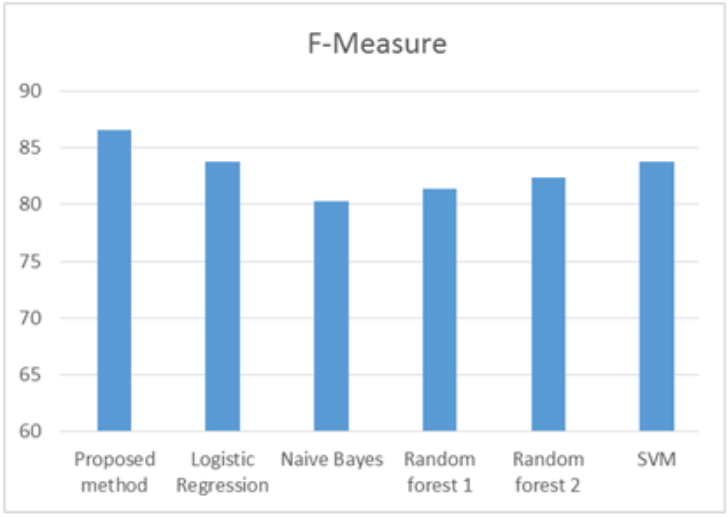

Fig. 7. Comparison of various models according to f-value criteria

Table (8) presents the results obtained from the proposed system's implementation and the basic classifications. Among the compared methods, no significant superiority was observed. The Logistic Regression method outperformed other models in terms of accuracy and recall, surpassing the SVM method (the best second-best method) by 0.67 in the mentioned criteria. In terms of Precision and F-value, the SVM method exhibited the best performance, with a difference of 1.85 and 0.03, respectively, compared to the Logistic Regression method (the second-best model). As indicated in Table 8, the proposed method achieved values of 85.57 for accuracy, 86.57 for recall, 84.51 for precision, and 86.54 for F-value. Comparing the results recorded in this table reveals that the proposed method outperformed the basic classifications.

Table 8. The values achieved for Accuracy, recall, precision, and F-value criteria by applying the proposed system and basic classifications

Proposed method	Logistic Regression	Naive Bayes	Random forest 1	Random forest 2	SVM	Criterion
85.57	83.75	80.36	82.37	82.75	83.08	Accuracy
86.57	83.75	80.36	83.26	83.26	83.08	Recall
84.51	83.70	80.29	80.22	81.29	85.55	Precision
86.54	83.73	80.31	81.41	82.36	83.76	F-Measure

6. Future Works

Parameter optimization can be considered for future research. One of the solutions that can be considered is optimizing the parameters with the help of evolutionary methods. Evolutionary methods search the desired problem space with different strategies and try to optimize the investigated parameters. There are different types of evolutionary methods that can be used, including genetic methods, memetics, particle swarm optimization, harmony search, gray wolf, etc.

7. Conclusion

In this research, the textual data from the dataset underwent multiple pre-processing steps. Additionally, logistic regression, NB, two versions of random forest, and SVM techniques were examined in the continuation. The investigation results indicated that logistic regression demonstrated superiority in terms of accuracy and recall, while support vector machines showed higher precision and F-measure. After performing these tests, the results of the combination of methods were analyzed using the majority voting technique. In the ensemble method, based on the majority vote, the output of the majority of the classification methods is determined. The final prediction is based on the result of the majority of classifiers. The results of the tests showed that the use of the majority voting technique can significantly increase the obtained performance. So ,in the final result obtained, the Accuracy criterion was improved by 2.78 compared to the previous best value.

There are some limitations in this field of research; the most important one is the limited access to text data for diseases. Text data in hospitals is usually not collected carefully, and many text data are not stored in computer systems. If data store correctly, access to them usually is hard work. As we said earlier from the management point of view, it is very important to have a possibility to diagnose cancer from textual data. With access to a large amount of textual data collected about diseases, health system managers can propose an artificial intelligence-based mechanism for cancer diagnosis. This system can also be used as an assistant to a specialist to help a specialist physician in the field of diagnosis. Managers can also consider new methods of teaching medical students by using artificial intelligence-based solutions for disease detection. By using methods based on artificial intelligence, the education process can be carried out at a lower cost.

Reference

Aggarwal, C. C. & Zhai, C. (2012). Mining text data. Springer Science & Business Media. ISBN: 978-1-4614-3222-7 (Print) 978-1-4614-3223-4. (Online)

Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. "Fake news detection using machine learning ensemble methods." Complexity 2020 (2020).

Aldhaeebi, Maged A., Khawla Alzoubi, Thamer S. Almoneef, Saeed M. Bamatraf, Hussein Attia, and Omar M. Ramahi. "Review of microwaves techniques for breast cancer detection." *Sensors* 20, no. 8 (2020): 2390.

Asgharizadeh, E., Kadivar, M., Noroozi, M., Mottaghi, V., Mohammadi, H., & Chobar, A. P. (2022). The intelligent traffic management system for emergency medical service station location and allocation of ambulances. *Computational intelligence and neuroscience*, *2022*.

Bhatia, Siddharth, Yash Sinha, and Lavika Goel. "Lung cancer detection: a deep learning approach." In Soft Computing for Problem Solving, pp. 699-705. Springer, Singapore, 2019.

Botlagunta, Mahendran, et al. "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms." Scientific Reports 13.1 (2023): 485.

Braiman L, Random forests, Machine Learn 2001; 45: p. 5-32.

Chandra, Tej Bahadur, Kesari Verma, Bikesh Kumar Singh, Deepak Jain, and Satyabhuwan Singh Netam. "Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble." Expert systems with applications 165 (2021): 113909.

Chobar, A. P., Adibi, M. A., & Kazemi, A. (2022). Multi-objective hub-spoke network design of perishable tourism products using combination machine learning and meta-heuristic algorithms. *Environment, Development and Sustainability*, 1-28.

Conceição, Sofia IR, and Francisco M. Couto. "Text Mining for Building Biomedical Networks Using Cancer as a Case Study." Biomolecules 11, no. 10 (2021): 1430.

Dildar, Mehwish, Shumaila Akram, Muhammad Irfan, Hikmat Ullah Khan, Muhammad Ramzan, Abdur Rehman Mahmood, Soliman Ayed Alsaiari, Abdul Hakeem M. Saeed, Mohammed Olaythah Alraddadi, and Mater Hussen Mahnashi. "Skin cancer detection: a review using deep learning techniques." *International journal of environmental research and public health* 18, no. 10 (2021): 5479.

Erdem, Ebru, and Ferhat Bozkurt. "A comparison of various supervised machine learning techniques for prostate cancer prediction." Avrupa Bilim ve Teknoloji Dergisi 21 (2021): 610-620.

Garg, Rishu, Saumil Maheshwari, and Anupam Shukla. "Decision support system for detection and classification of skin cancer using CNN." In Innovations in Computational Intelligence and Computer Vision, pp. 578-586. Springer, Singapore, 2021.

Harkema, Henk, Wendy W. Chapman, Melissa Saul, Evan S. Dellon, Robert E. Schoen, and Ateev Mehrotra. "Developing a natural language processing application for measuring the quality of colonoscopy procedures." Journal of the American Medical Informatics Association 18, no. Supplement_1 (2011): i150-i156.

Hekler, Achim, Jochen S. Utikal, Alexander H. Enk, Axel Hauschild, Michael Weichenthal, Roman C. Maron, Carola Berking et al. "Superior skin cancer classification by the combination of human and artificial intelligence." European Journal of Cancer 120 (2019): 114-121.

Hjaltelin, Jessica Xin, et al. "Pancreatic cancer symptom trajectories from Danish registry data and free text in electronic health records." medRxiv (2023): 2023-02.

Hosseini, S., Ahmadi Choukolaei, H., Ghasemi, P., Dardaei-beiragh, H., Sherafatianfini, S., & Pourghader Chobar, A. (2022). Evaluating the performance of emergency centers during coronavirus epidemic using multi-criteria decision-making methods (case study: sari city). *Discrete Dynamics in Nature and Society*, *2022*.

https://scikit-learn.org/

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis− A survey. *Pattern Recognition*, *83*, 134-149.

Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H. and Chen, Y.J., 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy, 8.

Iqbal, Saqib, Ghazanfar Farooq Siddiqui, Amjad Rehman, Lal Hussain, Tanzila Saba, Usman Tariq, and Adeel Ahmed Abbasi. "Prostate cancer detection using deep learning and traditional techniques." IEEE Access 9 (2021): 27085-27100.

Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In 2007 international conference on convergence information technology (ICCIT 2007) (pp. 1541-1546). IEEE.

Jahangiri, S., Abolghasemian, M., Ghasemi, P., & Chobar, A. P. (2023). Simulation-based optimisation: analysis of the emergency department resources under COVID-19 conditions. *International journal of industrial and systems engineering*, *43*(1), 1-19.

Jahangiri, S., Abolghasemian, M., Pourghader Chobar, A., Nadaffard, A., & Mottaghi, V. (2021). Ranking of key resources in the humanitarian supply chain in the emergency department of iranian hospital: a real case study in COVID-19 conditions. *Journal of applied research on industrial engineering*, *8*(Special Issue), 1-10.

Kaur, Ramandeep, and Navdeep Kaur. "Improved Skin Cancer Detection Classification Residual Network Feature Engineering." In *2021 International Conference on* Computational Performance Evaluation (ComPE), pp. 671-675. IEEE, 2021.

Khorshid, Shler Farhad, Adnan Mohsin Abdulazeez, and Amira Bibo Sallow. "A comparative analysis and predicting for breast cancer detection based on data mining models." Asian Journal of Research in Computer Science (2021): 45-59.

L.A. Torre, F. Bray, R.L. Siegel, J. Ferlay, J. Lortet–Tieulent, A. Jemal, Global cancer statistics, 2012, CA, Cancer J. Clin. 65 (2015) 87–108.

Liew, Xin Yu, Nazia Hameed, and Jeremie Clos. "An investigation of XGBoost-based algorithm for breast cancer classification." Machine Learning with Applications 6 (2021): 100154.

Lisboa, Paulo JG, Alfredo Vellido, Roberto Tagliaferri, Francesco Napolitano, Michele Ceccarelli, José D. Martín-Guerrero, and Elia Biganzoli. "Data mining in cancer research [application notes]." *IEEE computational intelligence magazine* 5, no. 1 (2010): 14-18.

Mahesh, T. R., V. Vinoth Kumar, V. Muthukumaran, H. K. Shashikala, B. Swapna, and Suresh Guluwadi. "Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer." Journal of Sensors (2022).

Mojrian, Sanaz, Gergo Pinter, Javad Hassannataj Joloudari, Imre Felde, Akos Szabo-Gali, Laszlo Nadai, and Amir Mosavi. "Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system." In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-7. IEEE, 2020.

Mushtaq, Zohaib, Akbari Yaqub, Shaima Sani, and Adnan Khalid. "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets." Journal of the Chinese Institute of Engineers 43, no. 1 (2020): 80-92.

Nanglia, Pankaj, Sumit Kumar, Aparna N. Mahajan, Paramjit Singh, and Davinder Rathee. "A hybrid algorithm for lung cancer classification using SVM and Neural Networks." *ICT Express* 7, no. 3 (2021): 335-341.

Patel, Falguni N., Hitesh B. Shah, and Shishir Shah. "A Technique to Find Out Low Frequency Rare Words in Medical Cancer Text Document Classification." In Advances in Data Computing, Communication and Security: Proceedings of I3CS2021, pp. 121-132. Singapore: Springer Nature Singapore, 2022.

Pradhan, A. (2012). Support vector machine-a survey. International Journal of Emerging Technology and Advanced Engineering, 2(8), 82-85.

Ramasubramanian, C., and R. Ramya. "Effective pre-processing activities in text mining using improved porter's stemming algorithm." International Journal of Advanced Research in Computer and Communication Engineering 2.12 (2013): 4536-4538.

Riquelme, Diego, and Moulay A. Akhloufi. "Deep learning for lung cancer nodules detection and classification in CT scans." *Ai* 1, no. 1 (2020): 28-67.

Saadi, Mesgari, and Ranjbar Abolfazl. "Analysis and estimation of deforestation using satellite imagery and GIS." GIS Application in Environment, GISDevelopment. net (2000).

Soffer, Shelly, Eyal Klang, Noam Tau, Roni Zemet, Shomron Ben-Horin, Yiftach Barash, and Uri Kopylov. "Evolution of colorectal cancer screening research in the past 25 years: text-mining analysis of publication trends and topics." Therapeutic Advances in Gastroenterology 13 (2020): 1756284820941153.

Upadhyay, Darshana, Jaume Manero, Marzia Zaman, and Srinivas Sampalli. "Intrusion detection in SCADA based power grids: Recursive feature elimination model with majority vote ensemble algorithm." IEEE Transactions on Network Science and Engineering 8, no. 3 (2021): 2559-2574.

Weiss, S. M., Indurkhya, N. & Zhang, T. (2010). Fundamentals of predictive text mining: Springer Science & Business Media.

Ye, Z., Tafti, A. P., He, K. Y., Wang, K., & He, M. M. (2016). Sparktext: Biomedical text mining on big data framework. PloS one, 11(9), e0162721.

Zhu, X., Yao, J. and Huang, J., 2016, December. Deep convolutional neural network for survival analysis with pathological images. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 544-547). IEEE.