DOI: 10.24996/ijs.2022.63.11.34

ISSN: 0067-2904

Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers

Fatin Kadhim Nasser*, Suhad Faisal Behadili

Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 2/10/2021 Accepted: 25/1/2022 Published: 30/11/2022

Abstract

Data mining has the most important role in healthcare for discovering hidden relationships in big datasets, especially in breast cancer diagnostics, which is the most popular cause of death in the world. In this paper two algorithms are applied that are decision tree and K-Nearest Neighbour for diagnosing Breast Cancer Grad in order to reduce its risk on patients. In decision tree with feature selection, the Gini index gives an accuracy of %87.83, while with entropy, the feature selection gives an accuracy of %86.77. In both cases, Age appeared as the most effective parameter, particularly when Age<49.5. Whereas Ki67 appeared as a second effective parameter. Furthermore, K- Nearest Neighbor is based on the minimum error rate, and the test maximum accuracy for K_value selection with an accuracy of 86.24%. Where the distance metric has been assigned using the Euclidean approach. From previous models, it seems that Breast Cancer Grade2 is the most prevalent type. For the future perspective, a comparative study could be performed to compare the supervised and unsupervised data mining algorithms.

Keyword: Breast Cancer, Gini index, Entropy, confusion matrix, classification report.

الكشف عن سرطان الثدي باستعمال مصنفات شجرة القرار والجيران الاقرب

فاتن كاظم ناصر* , سهاد فيصل شيحان

قسم علم الحاسوب ، كلية العلوم ، جامعة بغداد، بغداد، العراق

الخلاصة

ان لتعدين البيانات دور كبير في مجال الطب واكتشاف الامراض وعلاجها وبالأخص سرطان الثدي الذي يعد من أخطر الامراض التي تصيب النساء والرجال حيث يعد من أكثر الامراض المسببة للوفاة. في هذا البحث تم استعمال خوارزميتين من خوارزميات تعدين البيانات, الا وهما خوارزمية Decision Tree (DT) وخوارزمية K Nearest Neighbor (KNN) الاقرب لتحديد درجة سرطان الثدي. حيث ان دقة DT مع باستعمال ال Gini index لاختيار العامل الاكثر تأثير في المرض تساوي %87.83 وان العامل المؤثر الاول في الاصابة هو العمر وبعده عامل Ki67. بينما في خوارزمية KNN ان مقدار الدقة يساوي %86.24. ومع كلتا الخوارزميتين ان الدرجة الثانية لسرطان الثدي هي الاكثر انتشارا.

*Email: fatin.nasser1201@sc.uobaghdad

1. **Introduction**

Breast cancer (BC) is increasing cell size and dividing it out of control. It can infect both males and females and is the most common disease-causing death among women in the world. Early testing can increase protection from disease and make treatment more useful. In spite of trying to develop treatment but advance BC still hard goals of therapy range from symptom palliation to extending survival [1][2]. Many reasons increase the risk of BC, like food, age, menarche, menopause, inheritance, and the number of children [3]. The diagnosis and prognosis of BC take a great deal of time for researchers. Machine learning (ML) and data mining (DM) play a great role in the detection and prediction of BC [4], such that DM analyzes a large amount of data and can help in the early detection of BC [5]. In this research, two classification algorithms were used: Decision Tree (DT) and $K - Nearest$ Neighbor (KNN) applied to the data to provide efficient detection of BC grade. In the first step, data is loaded and then preprocessed by removing noise data (clearing data) and treating missing data[6][7][8]. The most suitable method for treating missing data is imputation [9]. In this research, most frequently, strategy is used instead of mean strategy in order to keep the reality of data. After that, the data was divided into two sets: the train set and the test set, and then DM algorithms were applied and the results were combined to get the final result.

The remaining parts of this research are organized as follows: Section 2 demonstrated the related works that have been done in this field. Section 3 explains the theoretical parts of the ML techniques used in this research. Section 4 performs analysis of data and preprocesses it, which deals with noise in data and missing data. Meanwhile, section 5 process steps involve three phases: splitting data, building a model, and evaluating the model.

2. **Related Work**

DM plays a great role in many areas like physiological data [10][11][12], gene/protein position dataset prediction, molecular bioactivity estimation for drug development [13][14], the colon cancer and leukemia dataset [15] etc. Machine learning techniques are applied to multiple medical fields to enhance medical decision-making. In bioinformatics science, DM has important studies in the cancer area [16][15]. For example, using DT and KNN technique to detect BC [17]. Open source data is available for BC like Surveillance Epidemiology and End Results (SEER) instances [18], Wisconsin Breast Cancer (WBC) datasets [19], and Wisconsin Breast Cancer Diagnosis (WBCD) datasets [20]. This research concerns BC dataset, which was obtained from the Medical City Center in Baghdad. Different research in BC applied several DM techniques and uses different BC dataset resources as illustrated next. In [19], DT, KNN, and Naïve Bayes (NB) algorithms were used. The results show that DT gives best result than other two algorithms such that DT $accuracy = 93.18\%$. In [17], the KNN algorithm was applied to the WBC dataset for prediction BC with $accuracy = 94.35\%$. Furthermore, in [20], KNN and DT were applied to the WBCD dataset to diagnose whether BC is malignant or benign. Results show that the KNN classifier is the more competent ML algorithm when compared with the Decision-Tree classifier. Additionally, in [21], KNN, and Support Vector Machine (SVM) were applied to the WBCD dataset to identify BC. Both algorithms give good result: KNN $accuracy = 92.31\%$, and SVM $accuracy = 95.65\%$.

3. **Dataset Description**

In this research, the work based on a dataset containing nine biomarkers, which are $Ki67,\ HER/2\ , PR, ER, DX, Grade, Stage, Type$of surgery$, and\ Age$ taken from 940 patients suffering Breast Cancer (BC), study was designed in Oncology Teaching Hospital, Medical City, Baghdad, Iraq, in 2014-2016, data written by hand and suffering from missing data, data inconsistency, and noise data (outlier data). A description of these datasets is illustrated in Table 2.

Table 2: Dataset Description

Biomarker	Description
$\boldsymbol{Ki67}$	Represent cells protein which increases when cells break down into new cells. Tumor in cells can be measured by staining processes that are more positive for $Ki-67$, when $Ki-67$ reached 0.14 it becomes a deadline.
$\boldsymbol{HER/2}$	Human epidermal growth factor receptor 2 ($HER2$) are proteins that reside on top side of breast cells, when HER2 proteins increase can lead to a certain type of breast cancer. If $HER2-negative$ no breast cancer is found, else if $HER2-positive$ means breast cancer found. Staining depends on the value of HER2, when $HER/2=1$ staining is weak, $HER/2=2$ moderate stain with non complete staining of the whole-cell membrane and it is equivocal and $HER/2=3$ stain is strong.
$\boldsymbol{PR\&ER}$	Estrogen receptors (ER) and progesterone receptors (PR) cancer cells with these receptors depend on estrogen and related hormones, such as progesterone, to grow. If breast cancer cells have estrogen receptors, the cancer is called $ER-positive$ breast cancer. If breast cancer cells have progesterone receptors, the cancer is called $PR-positive$ breast cancer. If the cells do not have either of these 2 receptors, the cancer is called $ER/PR-negative$. About two-thirds of breast cancers are $ER\ and/or\ PR\ positive$.
$\boldsymbol{Stage}$	Staging is a way for explaining breast cancer spreading, involving the size of the tumor, whether it reaches to lymph nodes, if it reaches distant parts of the body, and what its biomarkers are. Staging can be done either before or after a patient undergoes surgery.
$\boldsymbol{Dx}$	As a diagnosis disease, Breast cancer can be diagnosed through multiple tests, including a mammogram, ultrasound, MRI and biopsy.
$\boldsymbol{Grade}$	Describes the appearance of cancer cells and tissue and there are 3 grades of breast cancer, where $Grade1$ meaning that well-differentiated carcinoma, $Grade2$ moderatlity differentiated carcinoma, and $Grade3$ poorly differentiated carcinoma.
$\boldsymbol{Type\ of\ surger}$	There are 2 basic types of surgery to remove breast cancer Lumpectomy. The surgeon removes the breast tumor and a small rim of normal tissue around it. The rest of the breast remains intact and Mastectomy. The surgeon removes the entire breast. In many, but not all, cases this includes the nipple and areola.
$\boldsymbol{Age}$	The classification depends on age.

4. Theoretical Consideration

Sometimes, datasets may suffer from missing data. There are several types of missing data. Missing Completely at Random (MCAR) implies lost data distribution if there is no basement between lost data and its known values. In addition, missing at random (MAR), if lost values are based on known values and not on lost values themselves. As well as Not Missing at Random (NMAR), if lost data doesn't rely on known values or lost values [9]. Treating missing values is performed in the preprocessing phase.

Before starting to build a model data can be used in a helpful way by partitioning data (split data) into two groups called the train set and test set. The training set is useful in building models and attribute groups; it is useful in evaluating parameters, comparing models, and all other actions in order to get a final model. A test set is used as the final step for these actions to evaluate model execution [22]. Basically data is divided either by $K - fold$ cross validation, in which dividing data into $k - number$ of blocks is called $fold$, suppose $k = 5$. Figure 1 illustrates structure of $k - fold$ cross validation. This data type is not suitable for time series data, small data, and unbalanced data, or by using train-test-set which involves dividing input data as a ratio between the train set and the test set. After the data splitting model is built, there are several DM algorithms that are applied to the data in order to obtain results and detect grade the of cancer. In this research, two algorithms, DT and KNN applied.

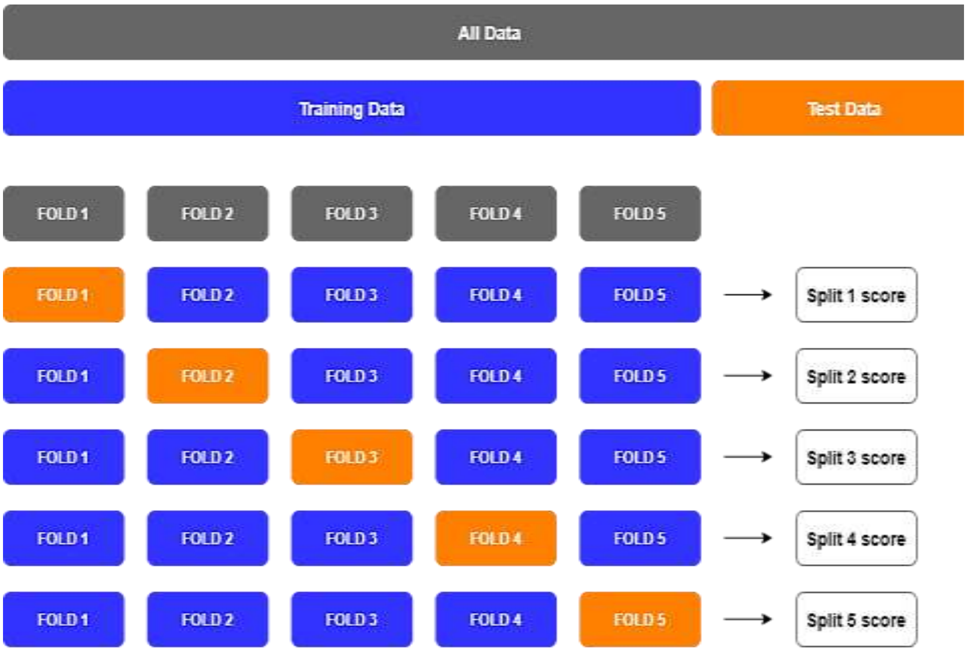

Figure 1: $k - fold$ cross-validation structure

DT is a non-parametric supervised algorithm which can be used for classification and regression and always for medical purposes. Structure of DT consists of multiple nodes, the one named root node. Connection between these nodes by edges such that each one has one incoming edge except the root. Some nodes have one or more outgoing edges called internal nodes or test nodes, and other nodes do not have an outgoing edge. This type is called decision nodes, also called leave or terminal [23][24]. In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attribute values, as shown in Figure 2.

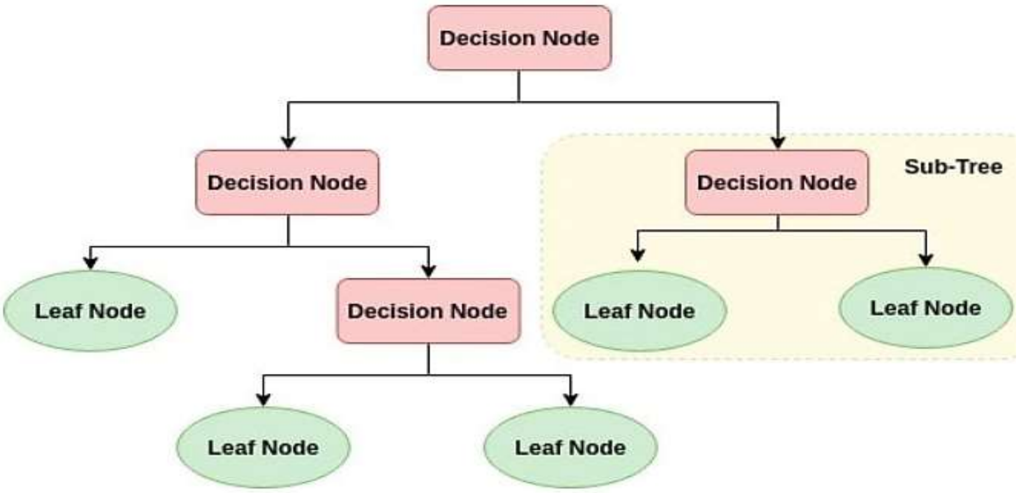

Figure 2: Decision Tree structure

Some criteria used to make decision, and which biomarker (attribute) are used as best partition for a train set and this biomarker used as testing in node of a tree, [25] these metrics are Gini index (GI) used to reduce classification error probability (misclassification) used in $CART$[26], it is computing in equation 1.

$$GI = 1 - \sum_{j}^{c} p_j^2 \tag{1}$$

Where, GI is Gini index, c is class labels, p_j is the probability class$i$. As well as, entropy is a randomness or impurity measurement in a system, entropy value is zero if all data is joined to one class, else entropy has a maximum value if the probability for each class is equal, entropy computed in equation 2.

$$Entropy = \sum_{i}^{c} -p_i * log_2(p_i) \tag{2}$$

Where, p_i is the probability of class i. Information Gain (IG) is dependent on entropy, whether a measurement of attribute can be helpful in classification or not. If an attribute is useful in classification, it causes increased IG value, and this attribute will be a good choice the splitting process, used by $ID3$ [26][27][28]. So, IG will be computed in equation 3 [29].

$$IG = entropy(parent) - average\ entropy(children) \tag{3}$$

KNN is a non-parametric supervised algorithm. This algorithm is famous because it is easy to understand. But it wastes memory because it stores the whole training set for classification and time complexity for test time. It saves all the training data and uses the whole training set for classification or prediction. This just saves all the values from the data set. It is important to define metrics for calculation distance between training set and test set [30], [31] Euclidean Distance defined as the square root of the sum of the squared differences between the two points of interest and it is favorite one by expert, the formula is in 2D space computed in equation 4 [32].

$$D_{(x_1,x_2)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4}$$

In addition, by using Manhattan distance, also called City Block distance, to calculate the distance between real vectors using the sum of their absolute difference, the formula is in 2D space, computed in equation 5.

$$D_{(x_1,x_2)} = |x_1 - x_2| + |y_1 - y_2| \tag{5}$$

As well as, by using Minkowski Distance, consider a generalization of Euclidean and Manhattan, and compute equation 6.

$$D = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{6}$$

Where, $p = 1$ is considered as Manhattan, $p = 2$ is considered as Euclidean. After building a model, it must measure the performance of the system. This is done by using a confusion matrix [33] as shown in Table 1 which gives information about system performance by storing True positive (TP) classifier predicts as positive and is true, False Positive (FP) classifier predicts as positive and is false, True Negative (TN) classifier predicts negative and is true, and False Negative (FN) classifier predicts negative and is false. From these values can compute $Precision$ which determines the ratio of number of instances assigned correctly as positive to classifier to total instances assigned as positive and computed in equation 7, and $Recall$ is ratio between number assigned correctly to class and all instances in the class and computed in equation 8. As well as $F1$ score defined as weighted average of $Recall$ and $Precision$, computed in equation.9, moreover, $accuracy$, which gives performance measurement of classifier, thus its computed in equation 10 [34][6], when imbalance dataset Recall and Precision are more useful than accuracy.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1\ score = \frac{2}{(\frac{1}{Precision} + \frac{1}{Recall})} \tag{9}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

Table 1: Confusion matrix structure

Predictive value	Actual value	
	TP	FP
	FN	TN

5. Materials And Methods

This step is divided into two phases: the preprocessing and processing phase to perform treatment on the original dataset and apply some DM algorithm to detect BC grade as shown later.

5.1. Preprocessing phase

This phase is a first step in treating data, [6] preprocessing involves several operations: data cleaning, data integration, data transformation, and data reduction. As shown in Figure 3, this process is often called data cleaning. [9] In order to deal with missing value either by ignore missing value or delete missing value or by using imputation technique, in this paper deal with imputation technique to fill missing value which contains several strategies such as mean**,** median, most frequent, and KNN, but practically mean strategy give outlier result, and most frequent strategy gives more suitable results in order to keep reality of data, so this strategy had been applied in this research.

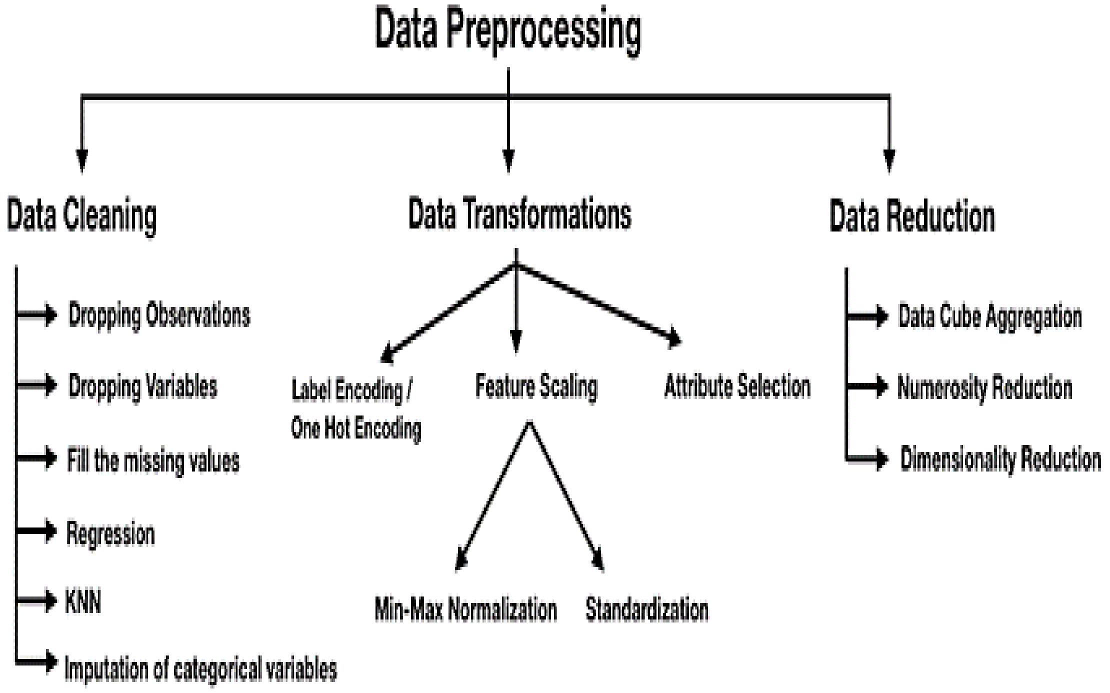

Figure 3: Preprocessing phase [26]

5.2. Processing phase

This step is the second phase in DM, which means applying DM algorithms on datasets in order to obtain effective results. This paper is concerned with two algorithms that are DT and KNN. At first, before building the classifier, the data must be split into a train set and a test set. In this paper, train set $size =\ 80\%$ of input data, test $size = 20\%$ of input data.

A. Decision Tree (DT) Algorithm

The observed dataset has been processed by using DT and the result explored in Figure 4 selection biomarker depends on GI criteria to predict the Grade of BC, such that there are three Grades: $Grade1,\ Grade2, Grade3$. Its $accuracy$: 81.48%, $Micro\ Precision$: 81.48%, $Micro\ Recall$: $81.48\ \%$ classification report shown in Table.3, and confusion matrix shown in Figure 5. The classification report gives more exact results than accuracy for classifier strength, easy interpretation, and discovery problems. Support defines the number of samples that belong to each class. However, it seems like a huge tree and difficult to analyze the result. So, the solution to this problem is by performing tree pruning with cost complexity to make tree more sharp features, decrease the size, and accuracy increased as shown in Figure 6 which represents pruning DT with GI. Also, its $accuracy$: 87.30 %, $Micro\ Precision$: 87.30 %, $Micro\ Recall$: 87.30 %, classification report shown in table 4, and confusion matrix shown in Figure 7. But if DT builds with entropy criterion as shown in Figure 8 with $accuracy$: 84.13%, $Micro\ Precision$: 84.13 %, $Micro\ Recall$: 84.13 %, confusion matrix shown in Figure 9, and classification report in Table 5. As well as, pruning DT can be performed with entropy criterion as shown in Figure 10, which has $accuracy$: 86.77%, $Micro\ Precision$: 86.77%, $Micro\ Recall$: 86.77% confusion matrix shown in Figure 11, and classification report in Table 6. The root node begins in Age biomarker all patients who have values less than 49.5 will be to the left of the nod, else will assign to the right. Then, left and right children's nodes are also split left child depends on $Ki67$ biomarker and right child on Age biomarker and the patient assigned to the sub tree depends on biomarkers value, this process repeated recursively and called recursive partitioning.

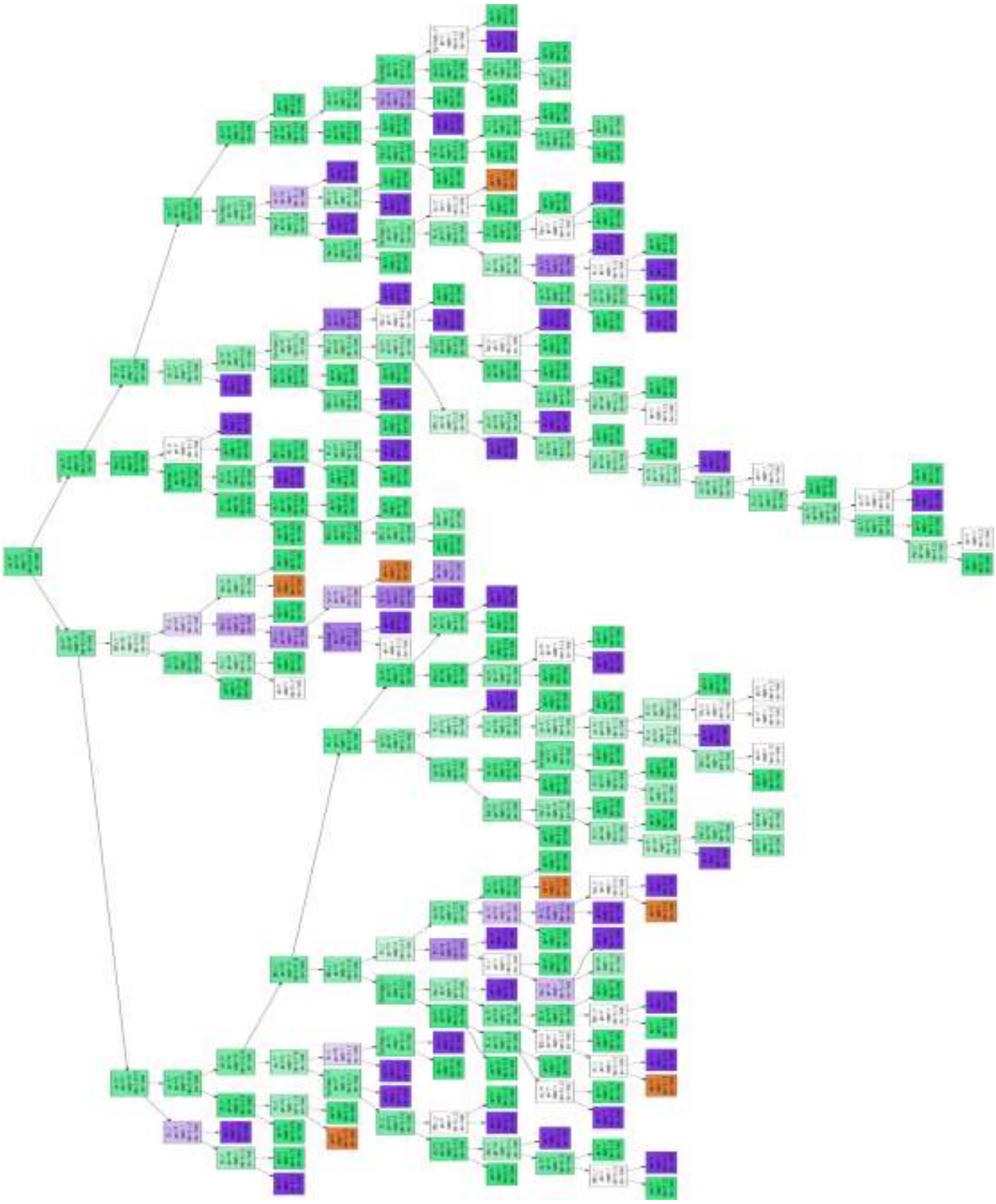

Figure 4: Decision Tree Result

Table 3: Decision Tree classification report

	Precision	Recall	f1-score	Support

Grade 1	0.00	0.00	0.00	4
Grade 2	0.90	0.91	0.90	165
Grade 3	0.21	0.20	0.21	20
Accuracy			0.81	189
Macro avg	0.37	0.37	0.37	189
Weighted avg	0.81	0.81	0.81	189

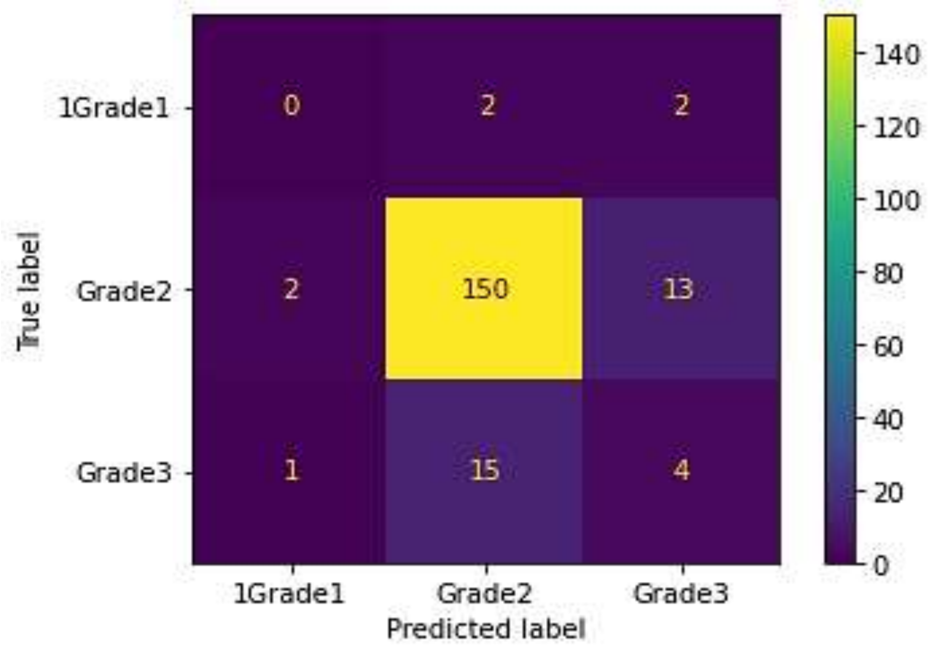

Figure 5: Decision Tree Confusion Matrix

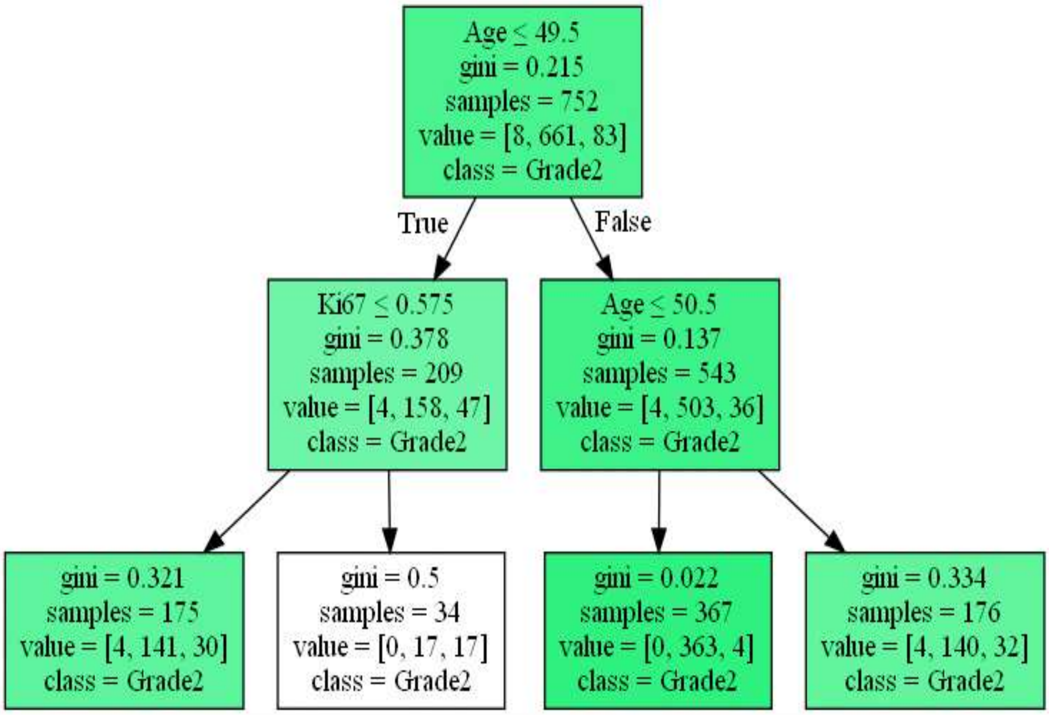

Figure 6: Decision Tree after pruning with *GI*

Table 4: Decision Tree classification report after pruning with Gini index

	Precision	Recall	F1-score	Support
Grade 1	0.00	0.00	0.00	4
Grade 2	0.87	1.00	0.93	165
Grade 3	0.00	0.00	0.00	20
Accuracy			0.87	189
Macro avg	0.29	0.33	0.31	189
Weighted avg	0.76	0.87	0.81	189

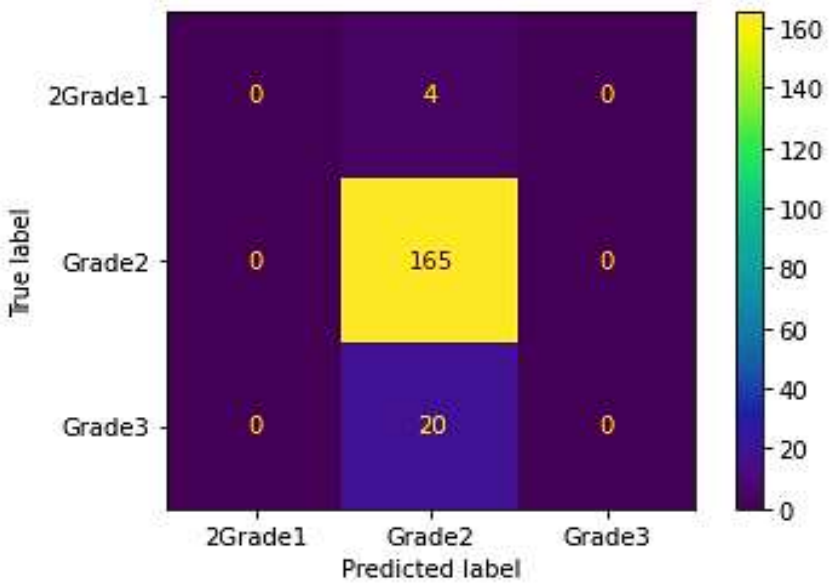

Figure 7: Decision Tree confusion matrix after pruning with GI

B. K-Nearest Neighbor (KNN) Algorithm

In this part, KNN is applied to the same observed dataset. First must determine value neighbor (k) take range from 1 to square root of number samples, to determine most appropriate k value which point has a minimum error and maximum accuracy as shown in Figure 12 and Figure 13 which illustrate ratio between $k - value$ and error, $k - value$ and accuracy respectively. From previous figures which showed error decrease when k is increased, and accuracy increase when k increased and point 13 is the optimal point. So, classifier $accuracy = 86.24\%, Micro\ Precision = 86.24\%, Micro\ Recall = 86.24\%$, whereas confusion matrix shown in Figure 14, and the classification report shown in Table 7.

Table 5: DT classification report with entropy before pruning

	Precision	Recall	F1-Score	Support
Grade 1	0.00	0.00	0.00	4
Grade 2	0.92	0.92	0.92	165
Grade 3	0.36	0.40	0.38	20
Accuracy			0.84	189
Macro avg	0.43	0.44	0.43	189
Weighted avg	0.84	0.84	0.84	189

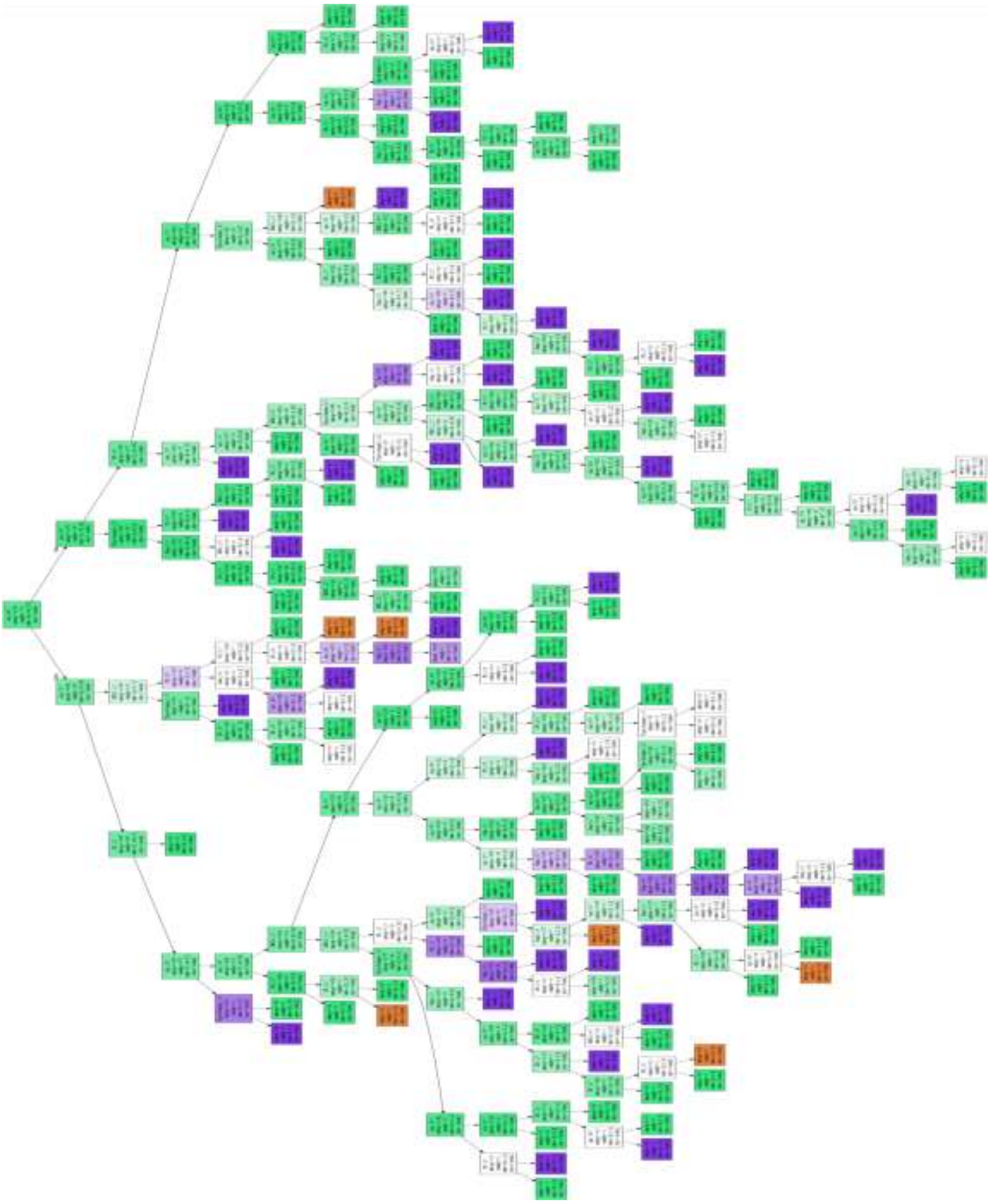

Figure 8: DT with entropy

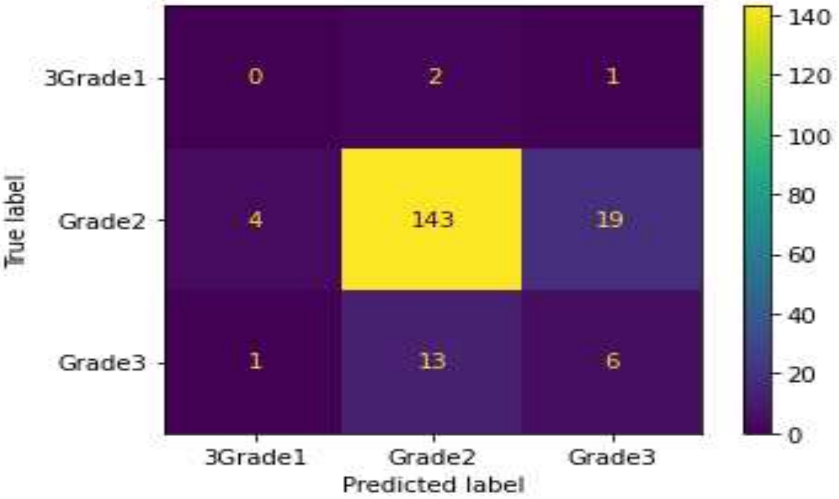

Figure 9: Confusion Matrix of DT with entropy

6. The Results Discussion

This research aims to analyze the performance of classification algorithms for BC based on the observed data, which are described in Table 2 and analyzed by KNN, pruning DT with GI, and pruning DT with entropy, depending on accuracy, Recall, and Precision value which shown in Table 8 seems that pruning DT with GI gives the best result from pruning DT with entropy, so it will depend in the classification process. A comparison between pruning DT with GI and KNN is carried out by this research. In DT Age biomarker is the most effective parameter and splitting process depending on its value, if $Age < 49.5$ then left child is checked depending on $Ki67$ biomarker, else the right child depending on Age biomarker as shown in Figure 6 and Figure 10. In KNN Euclidean approach for measuring distance between training object and test gives the best result than other approaches, the value of K selected in points with higher accuracy and lowest error as shown in Figure 12 and Figure 13.

This study implements KNN and DT techniques on an Iraqi real dataset, which produces results of less accuracy than the previous research [17, 19, 20, 21] that examined different datasets as presented in section 2. Furthermore, it confirms that DT has better results than KNN as in [19]. However, the used techniques are suitable for diagnosing BC disease as in [17, 19, 20, 21]. As well as, the results of DT have some limitations regarding their huge size, which confuses the observer when trying to analyze them. Thus, pruning with cost complexity has been used to overcome this problem. Moreover, this study confirmed that $Grade2$ is the dominant disease level depending on results of the classification report in Table 3 and Table 7.

7. Conclusions

In this study, we applied two different data mining (DM) classification techniques for breast cancer BC detection. The DM performance and accuracy were compared to evaluate the most effective algorithm in classifying the dataset. However, the Python language has been used as a supportive tool for analyzing the observed data with helpful libraries. Thus, from the results, it has been realized that pruning Decision Tree (DT) with Gini index (GI) gives better results with an accuracy of 87.30%, $precision\ of$ 87.30%, and $and\ recall\ of$ 87.30%, classification process depend on Age biomarker, where if $Age < 49.5$ left child depends on $Ki67$ value, else right child depends on Age. While K-Nearest Neighbor (KNN) gives accuracy 86.24%, $precision$ 86.24%, and $recall$ 86.24%, with $Euclidean$ gives the best result compared to other approaches. It concludes that both algorithms succeed in dealing with $accuracy, precision\ and\ recall$ metrics and it seems all results are equal. It is recommended to establish a collaboration bridge between BC medical centers and informatics scientists to find fruitful results. Finally, in the future, it is planned to increase the dataset sample to gain more model efficiency, which may produce more accurate results.

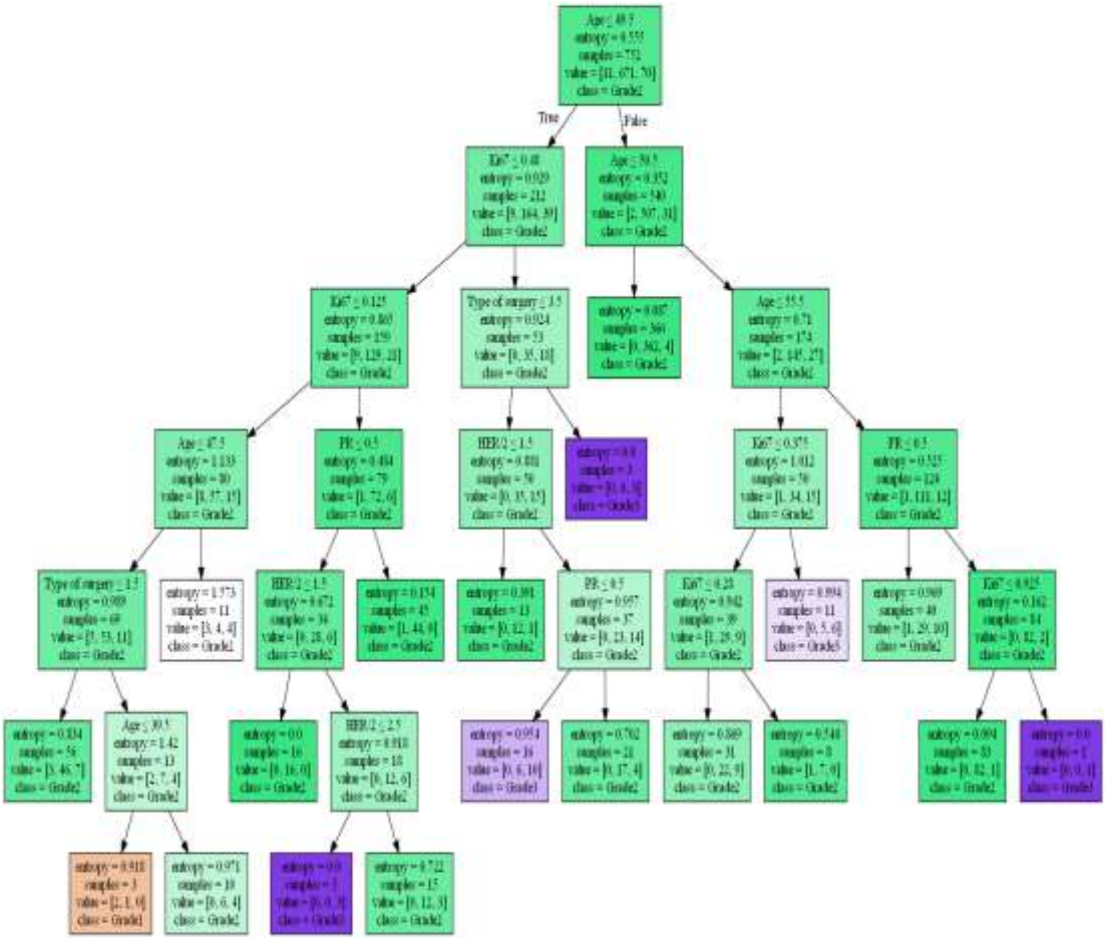

Figure 10: Pruning Decision Tree with entropy

Table 6: Pruning *DT* classification report with entropy

	Precision	Recall	F1-Score	Support
Grade 1	0.00	0.00	0.00	4
Grade 2	0.88	0.99	0.93	165
Grade 3	0.25	0.05	0.08	20
Accuracy			0.87	189
Macro Avg	0.38	0.35	0.34	189
Weighted Avg	0.80	0.87	0.82	189

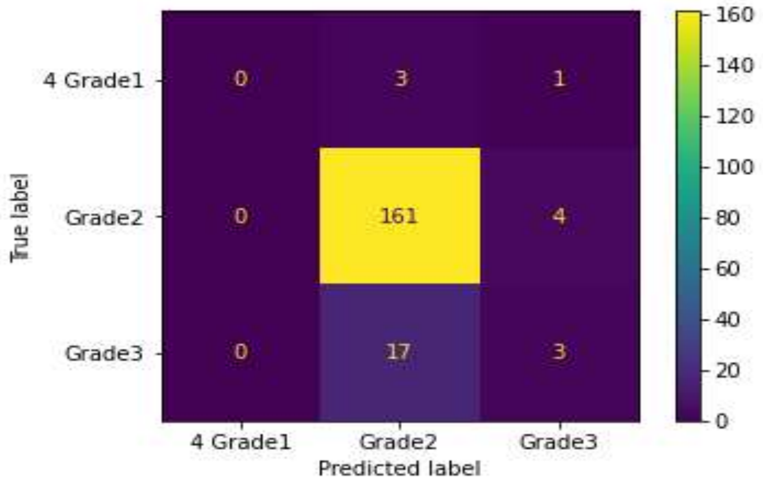

Figure 11: Confusion Matrix of Pruning DT with entropy

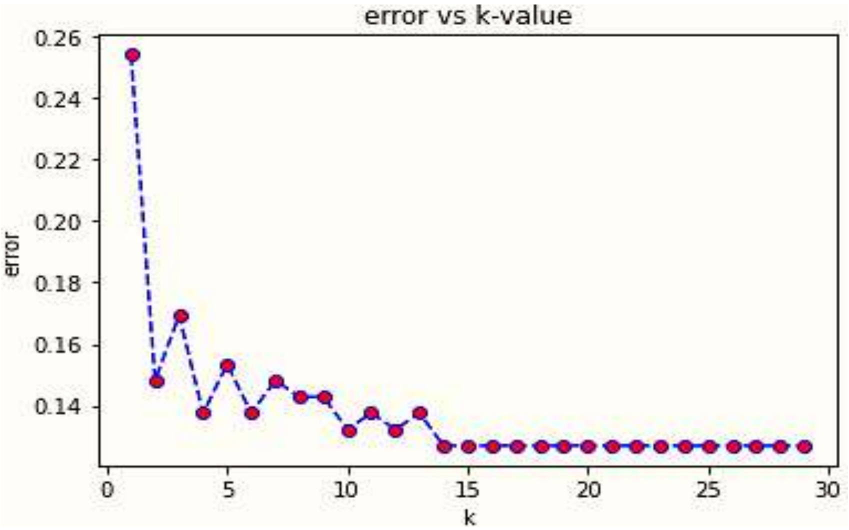

Figure 12: $K-value$ and error

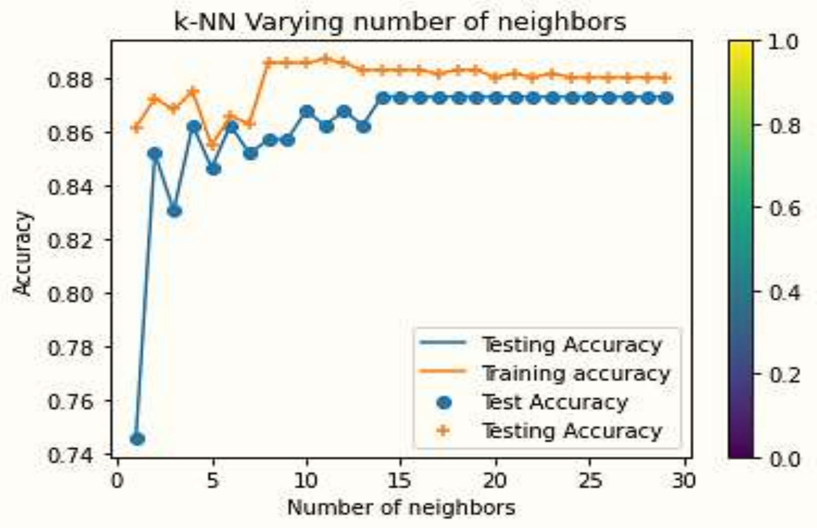

Figure 13: $k-value$ and accuracy

Table 7: KNN classification report

	Precision	Recall	F1-Score	Support
Grade 1	0.00	0.00	0.00	4
Grade 2	0.87	0.99	0.93	165
Grade 3	0.25	0.05	0.08	20
Accuracy			0.86	189
Macro Avg	0.29	0.33	0.31	189
Weighted Avg	0.76	0.86	0.81	189

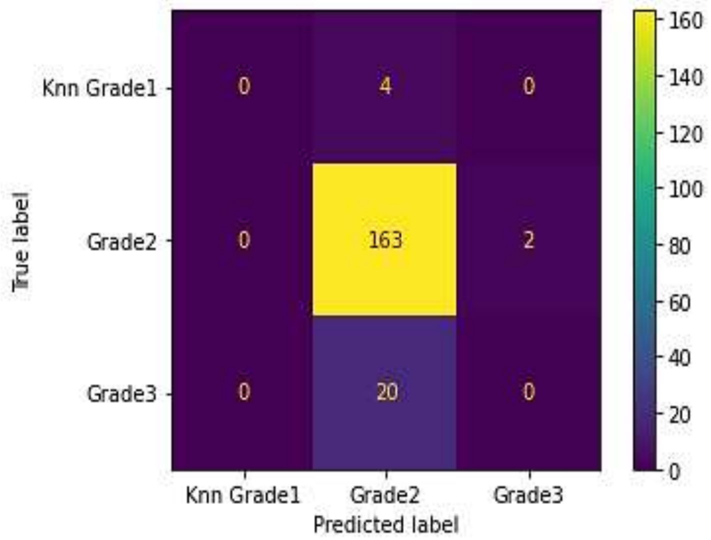

Figure 14: KNN Confusion Matrix

Table 8: Models accuracy

Model	Accuracy	Recall	Precision
Pruning ***DT*** with ***GI*** criterion	87.30 %	87.30 %	87.30 %
Pruning ***DT*** with entropy criterion	86.77%	86.77%	86.77%
KNN	86.24%	86.24%	86.24%

8. References

[1] H. Saad and N. Nagarur, “Data Mining Techniques in Predicting Breast Cancer,” *J. Appl. Sci.*, vol. 20, no. 3, pp. 124–133, 2020, doi: 10.3923/jas.2020.124.133.

[2] S. Faisal Behadili, M. S. Abd, I. Kamil Mohammed, and M. M. Al-Sayyid, “Breast Cancer Decisive Parameters for Iraqi Women via Data Mining Techniques,” no. May, 2019.

[3] N. Hamza Hassan and R. M. Ali, “Effect of life style (exercise and nutrition) on occurrence of breast cancer in women: A retrospective study in Babylon governorate,” *J Cont Med Sci*, vol. 1, no. 1, pp. 7–12, 2015.

[4] S. Gupta, D. Kumar, and A. Sharma, “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis,” *J. Comput. Sci.*, 2011.

[5] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, “Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets,” *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 43, no.

1, pp. 80–92, 2020, doi: 10.1080/02533839.2019.1676658.

[6] V. Brusic and J. Zeleznikow, "Knowledge discovery and data mining in biological databases," *Knowl. Eng. Rev.*, vol. 14, no. 3, pp. 257–277, 1999.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.

[8] F. Fatima, R. Talib, M. K. Hanif, and M. Awais, "A Paradigm-shifting from Domain-Driven Data Mining Frameworks to Process-based Domain-Driven Data Mining-Actionable Knowledge Discovery Framework," *IEEE Access,* 2020, doi: 10.1109/ACCESS.2020.3039111.

[9] A. Puri, S. Mata Vaishno, and M. Gupta, "Review on Missing Value Imputation Techniques in Data Mining," *Int. Conf. Mach. Learn. Comput. Intell.*, 2017.

[10] G. Getz, E. Levine, and E. Domany, "Coupled two-way clustering analysis of gene microarray data," *Proc. Natl. Acad. Sci. U. S. A.*, 2000, doi: 10.1073/pnas.210134797.

[11] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proc. 18th Int. Conf. Mach. Learn.*, 2001.

[12] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, "A generalized hidden Markov model for the recognition of human genes in DNA.," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1996.

[13] J. S. Eswari and C. Venkateswarlu, "Identification of Better Gene Expression Data for Mosquito Species Classification Using Radial Basis Function Network Methodology," *Open Bioinforma. J.*, vol. 11, no. 1, pp. 38–52, 2018, doi: 10.2174/1875036201811010038.

[14] S. N. Devi and S. P. Rajagopalan, "A Review on the Usefulness of Data Mining Techniques in Bio-Informatics," no. March, 2021.

[15] G. Krishna, B. Kumar, N. Orsu, and S. B., "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification," *Int. J. Adv. Res. Artif. Intell.*, 2013, doi: 10.14569/ijarai.2013.020508.

[16] M. Kaya Keleş, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Teh. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019, doi: 10.17559/TV-20180417102943.

[17] T. A. Assegie, "An optimized K-Nearest neighbor based breast cancer detection," *J. Robot. Control*, vol. 2, no. 3, pp. 115–118, 2021, doi: 10.18196/jrc.2363.

[18] N. Nazeer, B. Wajid, I. Nazir, and F. Gohar, "Prediction of Malignancy of Brain Cancer on SEER Dataset using Random Forest, SVM, and Naive Bayes Classifiers," *Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020*. 2020, doi: 10.1109/INMIC50486.2020.9318156.

[19] S. Uyun and L. Choridah, "Feature selection mammogram based on breast cancer mining," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 1, pp. 60–69, 2018, doi: 10.11591/ijece.v8i1.pp60-69.

[20] H. Rajaguru and S. R. Sannasi Chakravarthy, "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer," *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 12. pp. 3777–3781, 2019, doi: 10.31557/APJCP.2019.20.12.3777.

[21] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 226–229, 2018, doi: 10.1109/R10-HTC.2017.8288944.

[22] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Probl. Eng.*, vol. 2021, no. February, 2021, doi: 10.1155/2021/4832864.

[23] S. Kaur and H. Kaur, "Review of Decision Tree Data mining Algorithms: CART and C4.5," *Proceeding Int. Conf. Inf. Technol. Comput. Sci.*, vol. 8, no. 4, pp. 4–8, 2017.

[24] B. R. S. Kabra R R, "Performance Prediction of Engineering Students using Decision Trees," *Int. J. Comput. Appl. (0975 - 8887) Vol. 36- No.11, December 2011*, 2011.

[25] Q. N. Tran, "Using ANOVA to analyze modified Gini index decision tree classification," *Proc. 2008 Int. Conf. Data Mining, DMIN 2008*, no. January 2008, pp. 164–170, 2008.

[26] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.

[27] S. Lefkovits and L. Lefkovits, "Gabor Feature Selection Based on Information Gain," 2017, doi: 10.1016/j.proeng.2017.02.482.

[28] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[29] Linda Shapiro (University of Washington), "Information Gain Which test is more informative?," 2015, [Online]. Available: https://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf.

[30] S. Faisal, S. Sameer, I. Kamil Mohammed, and M. S Abd, "Review of medical diagnostics via data mining techniques," *Iraqi J. Sci.*, vol. 62, no. 7, pp. 2401–2424, 2021, doi: 10.24996/ijs.2021.62.7.30.

[31] V. Rodriguez, K. Sharma, and D. Walker, "Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements by," *Softw. Eng. Proj. (SWEN 670), Univ. Maryland, Univ. Coll. USA*, no. December 2018, 2018, doi: 10.13140/RG.2.2.20288.79361.

[32] A. M. Abdulazeez, "BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS : A REVIEW BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS : A," vol. 18, no. February, 2021.

[33] F. M. H. Fernandez and R. Ponnusamy, "A novel analysis and prediction of students' behaviour using semantic similarity-based improved J48 IL algorithm in personalized library ontology," *Int. J. Intell. Eng. Syst.*, vol. 11, no. 5, pp. 173–182, 2018, doi: 10.22266/IJIES2018.1031.16.

[34] D. Oyewola, D. Hakimi, K. Adeboye, and M. D. Shehu, "Using Five Machine Learning for Breast Cancer Biopsy Predictions Based on Mammographic Diagnosis," *Int. J. Eng. Technol. IJET*, vol. 2, no. 4, pp. 142–145, 2017, doi: 10.19072/ijet.280563.

[35] O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. Mostafa, and F. T. Al-Dhief, "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *Int. J. Eng. Technol.*, vol. 7, no. 4.36 Special Issue 36, pp. 160–166, 2018, doi: 10.14419/ijet.v7i4.36.23737.