Overcome Challenges in Primary Datasets: Practical Exploration and Preprocessing of Breast Cancer Dataset by Applying Machine Learning Algorithms

Samia M. Gharib[1], Nevine Makram Labib (Professor) [2], Yousri Rostom (Professor) [3], Salah Abdelmoneim (Professor) [4]

[1, 2] Computers and Information Systems Department, Sadat Academy for Management Sciences, Cairo, Egypt

3, 4 Oncology & Nuclear Medicine Department, Alexandria University, Alexandria, Egypt

sam4gharib@gmail.com, nevmakram@gmail.com , rostomy@hotmail.com , SALAH. ELMONEIM@alexmed.edu.eg

Abstract:

In 2020 Cancer caused 685,000 deaths worldwide, thus it is considered the second lethal disease globally. Because diagnosis of Metastatic Breast Cancer (MBC) patients is challenging, a prediction tool is needed during the diagnosis stage to define and prioritize patients who are more likely to develop metastasis and provide them with optimal palliative or supportive care. Machine Learning (ML) as a subset of Artificial Intelligence (AI) has been applied in oncology for early detection of cancer, identifying patients with high risk of survival, cancer morbidity, and mortality rate, besides predicting drug response. . One of the main applications of Machine Learning in public health is the identification and prediction of populations with high risk for developing specific adverse health outcomes, and development of appropriate targeted health interventions. Better data quality is crucial for better patient targeting and Informed Decision-Making. Also, the more and sufficient quality data the better machine learning model performance. Noisy or unclean may lead to inaccurate or faulty prediction, which is crucial in medical field. Consequently, data quality is essential for better Machine Learning model performance. The aim of this research is to determine the key challenges of using raw datasets, and illustrates how Machine Learning techniques can be used to explore and preprocess the dataset to overcome these challenges.

Keywords: Machine Learning - Data Exploration - Data Preprocessing - Breast Cancer Metastatic Dataset - Raw Dataset Challenges - Extra Trees Classifier - Random Forest Classifier

1. Introduction:

Breast cancer is the most prevailing cancer globally. In 2020, 2.3 million women were diagnosed with breast cancer which caused 685 000 deaths worldwide [1]. Metastasis is considered the principal cause of breast cancer-related mortalities. Only 5-10% of newly diagnosed breast cancer patients present with distant metastasis. While, nearly 30% of breast cancer patients diagnosed with early-stage disease are expected to develop metastasis, often months or even years later. The 5-year survival rate of patients with

localized disease is 99%, however it decreased greatly to 27% for patients with more aggressive distant disease [2].

Machine learning is a subset of artificial intelligence which applies a variety of statistical, probabilistic and optimization techniques to allow computers to learn from given historical examples in order to detect difficult to recognize patterns from large, noisy or complex data sets. Machine Learning techniques have been used for cancer prediction, recurrence and survivability [3].

Machine learning techniques are commonly used for detecting breast cancer in early stages. . There are two components of early detection: early diagnosis and screening. A prediction tool is needed to determine patients who are more likely to develop metastasis.

The feature selection, feature extraction and classification techniques are used to achieve promising results of the diagnosis and detection process. Thus Machine learning Prediction models help to decrease mortality rate among breast cancer diseases [4].

When implementing Machine Learning Models in the medical field, the following challenges and limitations should be taken into consideration[5]:

- Availability of data: Training of Machine learning models require large dataset in order to enhance performance and minimize errors.
- Quality of Data: Data quality is crucial in medical field. Intentional or unintentional errors during data entry may reduce quality of data. Preprocessing of data try to decrease the noise end errors of data to increase its quality.
- Highdimensionality of data: highdimensionality of primary datasets in healthcare increases the Machine Learning model complexity, learning time and cause overfitting. Overfitting means that the model doesn›t generalize well. Feature selection and extraction techniques can be used to reduce dimensionality of the datasets.
- Efficiency: Machine learning application in healthcare is essential in cases of high dimensionality datasets, difficulty of predicting parameters, need of time to predict correct results, inefficiency of normal methods for solving a problem.
- Privacy: During developing Machine Learning models in healthcare or medical domain, privacy issues has to be taken into consideration.

2. Methods & Materials

The data would be used for Metastasis Breast Cancer prediction model is a dataset of breast cancer patients provided by Department of Oncology & Nuclear Medicine, Alexandria University. It consists of 5236 patient›s records with 151 clinical and pathological variables. Machine Learning techniques were applied through python programming language and Anaconda Data Science Platform to explore and preprocess the dataset. During preprocessing Machine Learning techniques were used to Remove redundant Data & improve Column Titles, handle categorical values and scale data by python Libraries, handle missing values by K-Nearest Neighbor imputation, feature selection and resample data by Synthetic Minority Oversampling Technique (SMOTE) oversampling technique and Random Forest Classifier.

2.1 Dataset Description and Exploration

A primary dataset of Breast Cancer cases was provided by Clinical Oncology and Nuclear Medicine Department, Faculty of Medicine, Alexandria University. It consists of clinical records of 5236 instances for Female and Male patients diagnosed with breast cancer and 151 Attributes. Python libraries› been used to deal with the Dataset, and explore it. Using python Pandas library to explore the data, the following results were found:

1- The data set is multi-dimensional, small sized. So, dimensionality reduction MI techniques will be applied to handle this challenge.

2- There are redundant data like Serial, Data.No that would be deleted as not necessary for Machine Learning (ML) techniques that would be used and to hide the personal identifier of patients for ethics

integration.

3- The data frame has the following data types: object, float64, int64. Thus Preprocessing is essential to convert them to numeric values, as most of ML techniques work only with numeric values.

4- The class label would be used with Machine learning for Breast Cancer Metastasis (BCM) Prediction is Staging_groups, with 157 missing values. The class label is unbalanced, so ML techniques can be used to resample it.

non-metastatic 4747

metastatic 332

5- There are missing values in the data frame that exceed 90% in some columns.

6- The dataset contains demographic, clinical and therapeutic data about the patients. All the patients in the acquired dataset were diagnosed with breast cancer in different stages.

2.2 Data Preprocessing

Data preprocessing is the most important stage in the process of knowledge extraction from data. It can also improve performance of the ML models [6]. One of the most challenges faced when dealing with real world datasets is low quality data. Performing data analytics on poor-quality data, even applying the most powerful and optimal algorithms, may lead to inaccurate and unreliable results. Consequently, Data preprocessing before applying prediction process is inevitable to improve data quality. Moreover, data preprocessing is a stage in The Knowledge Discovery process which may that may require about 60% to 90% of the time necessary for knowledge discovery and contribute to 75% to 90% of the success of data mining cases [7].

The challenges found in some columns of the dataset in this study can be summarized in the following:

a- Columns labels that can make conflicts with python naming rules

b- Redundant data

c- Categorical values

d- Missing values

e- Feature Scaling

f- Dimensionality reduction

g- Unbalanced class label

Python programming language was used to handle column labels, remove redundant data as patient Id, serial, and convert categorical data to numeric. ML techniques was applied during the preprocessing stage to handle missing values. Missing values ratio exceeds 90% in some columns as shown in the following figure.

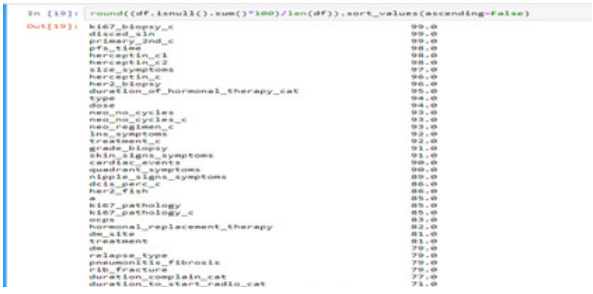

Figure1: - of missing values in the given dataset

2.2.1 Missing Values

Missing values in the dataset have a significant impact on the classification model performance, which can be summarized as follows [8]:

a-Reduce modeling efficiency.

b- Complexity of data preparation and analysis

c- Resulted in biased learning of the ML model.

d- Missing values causes may include but not limited to incorrect measurements, human error, and anonymous data.

Missing values may be expressed in the data as NaNs, blanks, undefined, or nulls. Messiness of data may be caused from improper and mistaken data entries, unavailability of data, problems of data collection, missing sequence, insufficient information, missing files, incomplete features [9]. Types of missing values can be; missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). It›s suitable to remove data with missing values in MAR and MCAR, but to bias can be generated when removing observation in MNAR. Thus carefulness is required

when removing missing values from the medical dataset [10]. Missing values in datasets can be overcome by many techniques including, removing records, or features and imputation. Imputation is e techniques for replacing missing value with specific estimated or potential value. Statistical and machine learning imputation techniques can be implemented such as Mean, Mode, regression, K-nearest neighbor, and ensemble algorithms[11]. An initial list of clinical and pathological diagnosing features was extracted from the dataset. And missing values was handled by using K-Nearest Neighbor imputation technique.

```
In [55]: df_new.isna().sum()
Out[55]: age_diag                      0
         marital_stat                  0
         menstrual_h                   0
         g                             0
         p                             0
         a                             0
         family_h_cancer               0
         medical_h                     0
         hormonal_replacement_therapy  0
         ocps                          0
         lump_symptoms                 0
         laterality_symptoms           0
         quadrant_symptoms             0
         nipple_signs_symptoms         0
         skin_signs_symptoms           0
         lns_symptoms                  0
         size_symptoms                 0
         laterality_mamogram           0
         quadrant_mamogram             0
         no_masses_mamogram            0
         size_mamogram                 0
         nipple_signs_mamogram         0
         skin_signs_mamogram           0
         lns_mamogram                  0
         biopsy_type                   0
         histologic_type_biopsy        0
         grade_biopsy                  0
         er_biopsy                     0
         pr_biopsy                     0
         her2_biopsy                   0
         ki67_biopsy                   0
         staging_groups                0
         dtype: int64
```

Figure 2: Percentage of missing values after imputing by KNN technique

2.2.2 Feature Selection

It›s the process of selecting most relevant features to improve the classification process. Feature selection methods are:

a- Filter Method: used to select the variables regardless the selected classification model. Select features by correlating the predictors and class label by collecting features most relevant to the classification.

b- Wrapper Method: used to select features by making combinations of features and finding interaction between them.

c- Embedded Methods: tries to integrate the advantages of wrapper and filter methods. Feature selection is carried out while executing the classification algorithm [12]. Extra trees, random forest classifiers were applied separately to select the most important features to reduce dimensionality. They both show the following results:

Feature selection Technique	Extra Trees Classifiers	Random Forest Classifier
Most Important Features Selected	Histologic_type_biopsy Age_diag Size_mamogram Size_symptoms Ki1V_biopsy Quadrant_mamogram Medical_h Quadrant_symptomsg	Histologic_type_biopsy Age_diag Size_mamogram Size_symptoms Ki1V_biopsy Quadrant_mamogram Medical_hg Quadrant_symptoms

Table 1: Most important features selected by Extra trees and Random forest classifiers

From the previous table we found that the resulted most important features of the 2 techniques are close , except the order of quadrant_syptoms and g.

2.2.3 Oversampling

Class label that ‹ll be used in Breast Cancer Metastasis prediction is unbalanced as most of the patients are non-metastatic.

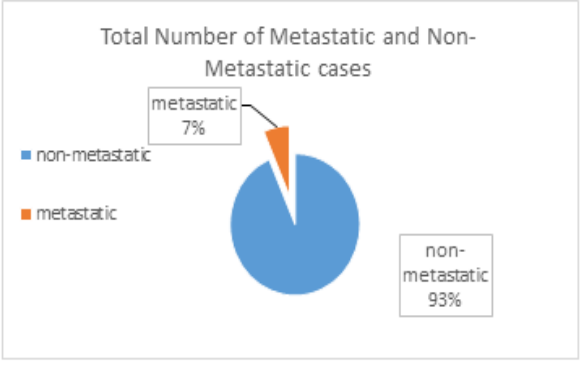

Figure 3: Percentage of metastatic and non-metastatic patients

Oversampling methods can be applied to handle unbalanced class label of the dataset. Oversampling can be applied by multiply the number of minority class members of the training

part after splitting the dataset. Oversampling reduce loss of information during training of ML model, as it tries to retain minority and majority observation of the class. Oversampling drawbacks may result from taking longer time for the model training and overfitting. Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique in which synthetic instances created to replicate the minority class and increase its number of instances in the training set. This is done by generating two key parameters which are the number of instances (n) and the nearest neighbors (k) [13].

When applying Decision tree classifier before oversampling the dataset, the following results were acquired:

```
In [441]: # Model Accuracy, how often is the classifier correct?
          print("Accuracy:",metrics.accuracy_score(y_test1, y_pred1))

          Accuracy: 0.9322200392927309

In [442]: report = classification_report(y_test1, y_pred1)
          print(report)

                        precision    recall  f1-score   support

                     0       0.93      1.00      0.96       951
                     1       0.00      0.00      0.00        67

              accuracy                           0.93      1018
             macro avg       0.47      0.50      0.48      1018
          weighted avg       0.87      0.93      0.90      1018
```

Figure 4: Evaluation Metrics of Decision tree classifier before oversampling

Applying SMOTE oversampling to the dataset we got the following evaluation results»

```
In [446]: # Model Accuracy, how often is the classifier correct?
          print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

          Accuracy: 0.6204724409448819

In [447]: #model = DecisionTreeClassifier()
          clf2 = DecisionTreeClassifier()

          # Train Decision Tree Classifer
          clf2 = clf.fit(x_train,y_train)

          #Predict the response for test dataset
          y_pred = clf.predict(x_test)

In [448]: report = classification_report(y_test, y_pred)
          print(report)

                        precision    recall  f1-score   support

                     0       0.68      0.43      0.52       932
                     1       0.59      0.81      0.68       973

              accuracy                           0.62      1905
             macro avg       0.64      0.62      0.60      1905
          weighted avg       0.64      0.62      0.61      1905
```

Figure 5: Evaluation Metrics of Decision tree classifier before oversampling

3. Machine learning Algorithms

The following ML algorithms was applied during the preprocessing stage to overcome the challenges of the dataset and prepare the data for the following stage of applying metastasis classification models.

3.1 K –Nearest Neighbour (KNN)

K-Nearest Neighbor (KNN) is an ML technique used to predict both discrete attributes and continuous features without building predictive model for each feature has missing values. Thus it can easily handle cases with multiple missing values. K-Nearest Neighbor (KNN) looks for the most similar instances, the algorithm searches through all the data set. This›s can be considered a critical limitation of this algorithm in large datasets analysis. The KNN imputation provide very good results for missing data imputations , even in cases the training sets had a large amount of missing data [14].

3.2 Random Forest Classifier (RFC)

It›s an ML supervised Learning method that can handle Classification and Regression problems. RFC consists of hundreds of Decision Trees. Each Decision Tree›s node performs a question about the data. The branches represent all possible answers to that question. RFC is an ML technique that combine a hundred decision trees. RFC is popular because of high accuracy and low computation costs of its performance [15].

3.3 Decision Tree (DT)

It›s a classification ML technique used to make prediction of categorical class names, classify knowledge on the basis of training sets and class labels, and to classify new data [16].

-A classification ML algorithm in which a set of

decision trees are generated from various subsets of the training data set. Averages used to check the accuracy of the prediction by estimating the model that fits the random decision tree [16].

3.5 SMOTE

Data is usually imbalanced in the medical field. In binary classification machine learning models, Imbalanced data means unequal distribution of the positive and negative classes. Imbalanced datasets may lead to a bias in the model and therefore, decrease the accuracy of the model to predict the minority class. To overcome imbalanced classes, various under-resampling, over-sampling, and hybrid methods can be used to improve the Machine learning model performance when predicting the minority class. On the other hand, these data sampling techniques have many disadvantages. While the oversampling techniques may increase the time or complexity of the model and also overfitting , the under-sampling techniques may cause loss of information.[17].

The Synthetic Minority over Sampling Technique (SMOTE) is a preprocessing technique for handling imbalanced datasets, by creating synthetic instances to oversample the minority class [18].

4. Conclusion

The main objective of this work is to present some results of data exploration and preprocessing by applying Machine Learning techniques on the given Breast Cancer patient›s dataset. In this work, the dataset was analyzed to determine the main challenges and the required ML techniques to overcome them. These challenges are redundant data, categorical data, missing values, unbalanced class label, and dimensionality reduction. Nine features were selected according to their importance by ML feature selection Decision trees and random forest models. Histologic_ type_biopsy, Age_diag, Size_mamogram, Size_ symptoms, Ki67_biopsy, Quadrant_mamogram, Medical_h, Quadrant_symptoms and g (Gravidity) which are the most important features suggested by ML to be used in Breast Cancer Metastasis prediction. As the dataset is small sized removal of missing values is not recommended in this case and instead ML Imputation techniques is used. K-Nearest Neighbor algorithm was used to impute multi- missing values of the dataset. Staging group, the class label to be used in Breast Cancer Metastasis prediction was determined. The class label was unbalanced, so Machine Learning oversampling was used to handle it by Synthetic Minority Over-sampling Technique (SMOTE). The dataset is ready for the next phase in which the Breast Cancer Metastasis prediction model is applied.

References

[1] "Breast cancer. World Health Organization." https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] S. Antar, N. Mokhtar, M. Adel, and A. K. Seleem, "Association of polymorphisms in metastasis suppressor genes NME1 and KISS1 with breast cancer development and metastasis," vol. 1, pp. 1-11, 2020.

[3] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," pp. 59-77, 2006.

[4] G. C. O. (GLOBOCAN), "Estimated number

of cancer cases in 2018, Egypt." [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/818-egypt-fact-sheets.pdf.

[5] A. M. Rahmani et al., "Machine learning (ML) in medicine: Review, applications, and challenges," Mathematics, vol. 9, no. 22. pp. 1-52, 2021, doi: 10.3390/math9222970.

[6] M. Zarei, A. Rezai, and S. S. Falahieh Hamidpour, "Breast cancer segmentation based on modified Gaussian mean shift algorithm for infrared thermal images," Comput. Methods Biomech. Biomed. Eng. Imaging Vis., vol. 9, no. 6, pp. 574-580, Nov. 2021, doi: 10.108021681163.2021.1897884/.

[7] Z. Sajjadnia, R. Khayami, and M. R. Moosavi, "Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services," Cancer Inform., vol. 19, pp. 7-12, 2020, doi: 10.11771176935120917955/.

[8] A. Elhassan, S. M. Abu-Soud, F. Alghanim, and W. Salameh, "ILA4: Overcoming missing values in machine learning datasets - An inductive learning approach," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.02.011.

[9] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010-2021)," Informatics Med. Unlocked, vol. 27, p. 100799, 2021, doi: 10.1016/j.imu.2021.100799.

[10] S. F. Huang and C. H. Cheng, "A safe-region imputation method for handling medical data with missing values," Symmetry (Basel)., vol. 12, no. 11, pp. 1-19, 2020, doi: 10.3390/sym12111792.

[11] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, A survey on missing data in machine learning, vol. 8, no. 1. Springer International Publishing, 2021.

[12] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim, "Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification," Innov. Smart Cities Appl. Ed. 2, 2019.

[13] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting breast cancer via supervised machine learning methods on class imbalanced data," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 54-63, 2020, doi: 10.14569/IJACSA.2020.0110808.

[14] G. E. A. P. A. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," Front. Artif. Intell. Appl., vol. 87, no. May 2014, pp. 251-260, 2002.

[15] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," IEEE Access, vol. 9, pp. 7869-7884, 2021, doi: 10.1109/ACCESS.2020.3047942.

[16] S. Ozdemir, Principles of Data Science. Packt Publishing Ltd, 2016.

[17] R. Gupta, R. Bhargava, and M. Jayabalan, "Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models," Proc. - Int. Conf. Dev. eSystems Eng. DeSE, vol. 2021-Decem, no. January, pp. 162-167, 2021, doi: 10.1109/DESE54285.2021.9719398.

[18] S. H. Abdulla, A. M. Sagheer, and H. Veisi, "Improving Breast Cancer Classification Using (SMOTE) Technique and Pectoral Muscle Removal in Mammographic Images," no. December, 2021, doi: 10.13164/mendel.2021.