ISSN 1022-8594
JUJS

Vol. 43, No. 1, pp. 63-78

A Study on Machine Learning Algorithms with Different Encoding Techniques for Identifying the Right One for Patients' Big Data

Subrata Kumar Das[a,*], Mohammad Zahidur Rahman[a]
[a]*Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh*

Abstract. In predictive modeling, categorical features often arise problems because most supervised machine learning algorithms can read numerical data as input instead of categorical attributes. So, many encoding techniques are used to convert categorical values into a machine-understandable format. Besides, different classifier algorithms could show their performance differently on the Big dataset. Therefore, the study goal is to find a learning model that will be a better-suited approach to a large volume of patients' data. This study also checks which encoding technique help to provide the high accuracy of the trained models. We applied here some encoding techniques on patients' data individually and their composite strategies to training machines. However, encoding techniques applied to categorical features and models learned as a classifier do not perform well and provide better performance. Some models trained here using various encoding techniques do not even work when facing the patients' Big data. Moreover, the training time of all machine learning models was not the same for the dataset. Therefore, this paper would help developers to choose reliable machine learning models to design their systems considering patients' Big data.

Keywords: Big Data, Encoding Techniques, Healthcare Data, Machine Learning Algorithms, Statistical Metrics.

1. Introduction

Healthcare organizations collect data, both structured and unstructured, about patients and their medication in electronic format. The stored data can help the physician making better clinical decisions by reviewing the patient's previous health conditions and medication. In this sector, different categorical data are required to store with patient records, such as gender, birth date, blood group, and more. Those categorical data are often analyzed and applied to characterize the patient’s status. However, making

* Author for correspondence e-mail: sdas_ce@yahoo.com

a decision on patients' Big data without computational techniques that can handle a large volume of datasets is not possible [1].

Therefore, machine learning is being important field in the medical sector for creating different classifications. For instance, to classify healthy and unhealthy patients, identify abnormalities, different diseases, treatment selection, etc. [2]. There are many algorithms to train the machine.

However, they all are not suitable to train a machine and get better performance because of the large volume of data. Moreover, the advancement of the medical sector in terms of electronic medical records has been remarkable, but the data they store is not much better than the traditional paper charts they replaced [3]. So, the electronic data need to be preprocessed and enhanced before making a decision using machine learning. The main problem is that most of the machine learning algorithms process only numerical inputs [4] [5] [6], so it is required to encode as the data contains both numerical and nominal values. There are many encoding techniques with their self advantages and disadvantages from different aspects to convert those categorical variables into numerical values. However, identifying the right encoding technique and machine learning algorithm could significantly impact the performance [7]. The selection of the right strategy and algorithm can provide low running time complexity and better efficiency.

Therefore, this article's aim is to apply various encoding techniques and composite encoding techniques to convert categorical variables to numerical variables for use in different machine learning algorithms. Besides, to evaluate the statistical metrics of the machine learning algorithms to identify which algorithm is suitable for patients' Big data.

The article organizes the remaining as follows. Section 2 presents significant related works. The detail of the dataset is provided in Section 3. Section 4 explains the methodology of the experiment. The ‘Results and Discussion’ section 5 points out the desired result and its finding. Finally, Section 6 includes a conclusion summarizing the work.

2. Related Works

Encoding techniques refereed to many terms are 'distributed representation' [7], 'entity embeddings' [8], 'dense encoding' [9], and simply 'encoding'. Potder et al. cited a comparative study of categorical variable encoding techniques [4]. This paper covered seven methods for encoding categorical variables UCI dataset [10] and learned only Artificial Neural Networks (ANN) model to test accuracy. Cerdat et al. presented categorical values encoding on the medical charges dataset to stress the significance of adapting encoding schemes to dirty categories [11]. Karthiga et al. [12] proposed a computer-aided diagnosis (CAD) system to diagnose automatic breast cancer using the one-hot encoding technique. An article was published to extract medically high-risk factors with machines in healthcare that enhanced accuracy using a categorical encoding technique [13]. A machine learning-based prognostic model was designed for giving early notification to individuals for COVID-19 infection [14]. The developed prognostic model used support vector regression and Random Forest classifier and converted categorical features using Label encoding. Sedighi et al. [15] proposed a two-stage data analytic framework (Stage I and Stage II) for classifying the survival and deceased statuses in Stage 1 and measuring the survival months for deceased females with cancer in Stage II. They use the One-hot technique to encode the categorical features in Stage I and Stage II. Mathur [16] wrote a chapter, 'How to Implement Machine Learning in Healthcare?' to mention the potential areas of the healthcare system. Mathur developed some machine learning models and used One-hot encoding to show the accuracy of the models. However, he used a small dataset with only 664 instances. A paper published on the machine learning method for heart disease data classification used One-hot encoding for data conversion [17].

Magolou et al. [18] introduced a computer model for determining the habits of life and the Health of the students of the Sultan Moulay Slimane University in Beni Mellal. They designed a text classification model based on the deep learning approach including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) by considering a semantic model with One-hot-encoding. Johannemann et al. [19] cited an article on lower-dimensional real-valued representations of categorical variables and

used the reference of the patient datasets from several hospitals. A general framework was designed by Nazábal to encode data automatically from different categorical features [20]. A comparison study was introduced by Arora et al. to seek the performance of different classification models for finding successful episodic memory encoding required from human stereo EEG subjects [21]. A two-layer model was proposed for patients' dementia for the early diagnosis using machine learning techniques [6]. In that study, various classification algorithms were learned to measure the performance of the model and compared them. Hancock and Khoshgoftaar [22] published a paper to show the classifier's performance in detecting fraud in Medicare data. They used different strategies for preprocessing datasets for classifiers and One-hot encoding for encoding categorical features. Ebrahimi et al. [23] cited an article on alcohol use disorder (AUD) prediction of alcohol to reduce the mortality caused by alcohol-related diseases. The study used a supervised machine learning model on a dataset from electronic health records (EHR). That study was performed on a small dataset of 2,571 patients only. Abdar et al. [24] learned the various classification algorithms for Coronary Artery disease and compared the performance of the models. Considering the performance of different models, the authors proposed a new machine learning methodology to detect Coronary Artery disease. Another machine learning algorithm was cited by Sujitha and Seenivasagam [25] to experiment with a combination of binary classification and multi-class classification for classifying nodules into malignant or benign nodules. They also used One-hot methods to encode the values for binary classification. Wang et al. tried to predict the prostate cancer patients die of non-cancer causes of death [26]. The authors used the Random forest (RF) learning model as well as the One-hot encoding technique. However, the accuracy of the model for predicting cancer patients was unclear. Another article was published by Hsieh et al. on pancreatic cancer with type 2 diabetes [27]. They applied logistic regression (LR) and artificial neural network (ANN) models to patients' datasets and found that the LR model predicted pancreatic cancer more accurately than the ANN model.

3. Source Dataset

For this study, we use healthcare data with categorical variables. The dataset is available in the openML database provided by Geoffrey Holmes et al. [28]. The source dataset consists of over 1000000 number of instances. The detail of the various features of the dataset is presented in Table 1. The raw dataset contains nine (9) input features and one (1) target feature with two label values.

Table 1. Feature names, data types and distinct values.

Features name	Data types	Unique values
Age	Nominal	9
Menopause	Nominal	3
Tumor-size	Nominal	12
Inv-nodes	Nominal	13
Node-caps	Nominal	2
Deg-malig	Nominal	3
Breast	Nominal	2
Breast-quad	Nominal	5
Irradiat	Nominal	2
Class (target)	Nominal	2

4. Methodology

To conduct the study, we preprocessed the data before learning the models. The steps we took into consideration are described below-

4.1 Data Preprocessing

The raw data may consist of missing values, so we first removed the missing values from the raw dataset. The categorical variables were then encoded using various encoding techniques and some composite encoding strategies, the mixed of more than one encoding strategy, detailed as follows. After encoding, the number of input features (10) changed in terms of dimensions shown in Table 2.

Table 2. Various encoding techniques and the dimension of input features.

Encoding techniques	Dimension after encoding
Binary Encoding	32
Frequency Encoding	10
Label Encoding	10
Mean Encoding	10
One-hot Encoding	41
Binary-Label Encoding	22
Frequency-Label Encoding	10
Frequency-One-hot Encoding	19
Label-Ordinal Encoding	10
Mean-Ordinal Encoding	10

Binary Encoding (BE): This encoding initially encodes categorical variables as integers and then converts them into binary code that is placed into separate columns. Let x be some values $x_1, x_2, \dots, x_n$ in a column. The categories are first replaced with numeric order starting from 1. The numeric data then are transformed into binary code.

Frequency Encoding (FE): Frequency Encoding counts the size of a category's occurrences in the dataset and converts them to a numerical value considering the total number of instances. Let x be some values $x_1, x_2, \dots, x_n$ in a column. Then the Frequency Encoding is measured in the following way.

$$FrequencyEncoding = \frac{FrequencyofUniqueValues, x_i}{TotalNumberofValues, \sum_{i=1}^{n} i}$$

Label Encoding (LE): The Label encoding algorithm encodes nominal data in an order. Let x be some unique text values $x_1, x_2, \dots, x_n$ in a column. Then the Label Encoding of the values is $x_i = i - 1$, where $i = 1, \dots, n$.

Mean Encoding (ME): This encoding replaces a categorical variable with the mean of the target feature. Let X be the 'categorical variable' and T the 'Target' variable. The categorical data are grouped based on the unique values from X, and their aggregated sum (S) is obtained over the T. The

aggregated count (C) of X is found over T. Then, the Mean Encoding is measured as follows.

$$MeanEncoding = \frac{S}{C}$$

One-hot Encoding (OhE): This encoding provides each level of the categorical feature with a fixed reference level [29]. One-hot Encoding converts a training feature with n instances and l distinct level values to l training attribute with n instances each. Each cell of the rows contains zero (0) or one (1) to indicate the absence or presence respectively. Let x be some unique text values $x_1, x_2, \ldots, x_n$ in a column. The one-hot Encoding of a specific value x_i is a vector v that contains zero for each component except for the i^{th} component that takes 1.

Some other composite encoding techniques, a combination of more than one encoding, were used to convert categorical variables based on the distinct value levels, the nature of the data if ordinal or not, etc. In this study, the used composite encoding techniques were Binary-Label Encoding (BL), Frequency-Label Encoding (FL), Frequency-One-hot Encoding (FO), Label-Ordinal Encoding (LOrd), and Mean-Ordinal Encoding (MOrd).

4.2 Learning Models

In this phase, the preprocessed dataset is used to train seven different machine learning models described below. The goal of learning the models is to predict the target class of the patient from the dataset and find out what models fit the patients' Big data efficiently for making a decision.

- Classification and Regression Tree (CART) : This algorithm predicts outcome variables' values based on other values. The output of a CART is a decision tree where each fork is a part of the predictor variable, and each end node consists of a prediction for the outcome variable. Let X denote the domain of x and Y the domain of y. If Y takes N individual values, the classifier could be expressed as a partition of X into N disjoint pieces such that $X = U_{n=1}^{N} A_n$, where $A_n = \{x: f(x) = n\}$. The regression tree is a constant or a relatively regression model that is fitted to the data in each partition.

- K-Nearest Neighbours (KNN): It can produce different outputs for classification from the same input features taking different values of K. Here, K indicates the number of closest neighbors that are considered for voting. Let x and y be some values $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ respectively. Then the distance is measured by the Euclidean method as follows.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Logistic Regression (LR): This classifier evaluates the weighted sum of the training features and uses the 'sigmoid' function on the weighted sum. The result from the sigmoid function can be represented as the probability of the positive classes in terms of binary classification. The model is learned by using weights during training. In mathematical terms, logistic regression is expressed for a training data point (x, y) as $P(Y = 1 \vee X = x) = f(x)$, where $f(x) = \frac{1}{1+e^{-x}}$.

- Linear Discriminant Analysis (LDA): This algorithm trains the model by projecting the higher dimension space features onto the lower dimension space. It evaluates between-class variance and within-class variance to generate lower-dimensional space [32].

- Naive Bayes (NB): It is a classifier based on the Bayes theorem, taking an assumption of independence among predictors. It is a collection of classification algorithms that share a common principle instead of a single algorithm. Let x and y be some values $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ respectively. Then the Naive Bayes approach is mathematically given by, $P(y \vee x_1, \dots, x_n) = \frac{P(x_1 \vee y)P(x_2 \vee y)\dots P(x_n \vee y)P(y)}{P(x_1)P(x_2)\dots P(x_n)}$

- Random Forest Classifier (RFC): The Forest classifier uses decision trees to classify. In this classifier, each tree splits nodes taking random features instead of the best features. The final resulted class

receives the majority vote individually from the trees. The Random Forest strategy divides the training features randomly [33].

- Support Vector Machine (SVM): This supervised machine learning algorithm maps n number of input features into an n-dimensional space. Then the classification is done by finding the hyperplane that separates the two classes very well [34].

4.3 Statistical Metrics Evaluation

In this section, we calculated the statistical metric of the machine learning models alongside various encoding techniques. Using the only accuracy as a performance measure for medical datasets of its imbalanced class distribution and the large volume of data is not a good idea [30] [31]. So, we measured other statistical metrics (Standard Deviation, training time) as well as accuracy to see the fitness of different models to patients' datasets.

5. Results and Discussion

We trained different models using Scikit-learn, an open-source Python library. All experiments were executed on Google Colab through the browser. We used 10-fold cross-validation to generate the best hyperparameters for each learning model. Twenty percent (20%) of data were taken as the test data. Table 3 shows the results of the accuracy of each training model received by taking various input features generated using different encoding techniques. From the Table, it is noticeable that KNN and SVM models did not run for the dataset. This result indicates that those two models (KNN and SVM) would not work for a large volume of data. SVM could not perform because the algorithm's training time depends on the dataset size and grows for a data point when it is infeasible to learn it. KNN could not act for big data for its lazy learning approach. KNN algorithm stores whole data and later make a decision only at run time. It finds the computation of distances for a selected point with all other points. So it takes a lot of processing time that might fail to work for a large dataset. The CART and RFC provide comparatively better accuracy than other algorithms, and both of them receive data from label encoding.

Table 3. Received accuracy from different encoding techniques.

Supervised Learning Algorithms	Accuracy(%)									
	Encoding Strategies					Composite Encoding Strategies				
	Binary Encoding	Frequency Encoding	Label Encoding	Mean Encoding	One-hot Encoding	Binary-Label Encoding	Frequency-Label Encoding	Frequency-One-hot Encoding	Label-Ordinal Encoding	Mean-Ordinal Encoding
Classification &Regression Trees	77.58	77.55	77.81	77.59	77.62	77.60	77.56	77.56	77.61	77.60
K Nearest Neighbor	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked
Logistic Regression	76.35	74.50	75.28	76.72	76.82	75.60	75.76	75.94	75.31	75.47
Linear Discriminant Analysis	76.39	74.09	74.95	76.51	76.54	75.37	75.32	75.54	75.00	75.18
Naive Bayes	71.45	70.42	72.19	73.72	73.16	71.92	72.03	71.08	72.24	72.36
Random Forest Classifier	77.93	77.88	78.04	77.94	77.99	77.92	77.87	77.83	77.90	77.90
Support Vector Machine	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked	Not worked

Another statistical metric, the standard deviation of various learning models, is presented in Table 4. The table shows that the two composite encoding techniques (Frequency-Label, and Label-Ordinal Encoding) and Label encoding perform with a lower standard deviation. In spite of having a lower standard deviation of the mixed encoding for LR and LDA than CART, the CART model performs better with the Label Encoding technique for the patient dataset due to having higher accuracy than LR and LDA.

Table 4. Comparison of standard deviation of different encoding techniques.

Supervised Learning Algorithms	Standard Deviation									
	Encoding Strategies					Composite Encoding Strategies				
	Binary Encoding	Frequency Encoding	Label Encoding	Mean Encoding	One-hot Encoding	Binary-Label Encoding	Frequency-Label Encoding	Frequency-One-hot Encoding	Label-Ordinal Encoding	Mean-Ordinal Encoding
Classification & Regression Trees	0.001712	0.001697	0.000868	0.001932	0.001563	0.001500	0.001743	0.001726	0.001524	0.001517
K Nearest Neighbor	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work
Logistic Regression	0.001054	0.001028	0.001394	0.001455	0.001324	0.001031	0.000696	0.001140	0.001076	0.001122
Linear Discriminant Analysis	0.001149	0.001293	0.001255	0.001157	0.001248	0.001128	0.000945	0.000901	0.000862	0.001013
Naive Bayes	0.001749	0.001697	0.001548	0.001581	0.001694	0.001645	0.001231	0.001030	0.001142	0.001181
Random Forest Classifier	0.001780	0.001756	0.001022	0.001858	0.001789	0.001651	0.001662	0.001920	0.001652	0.001845
Support Vector Machine	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work	Not work

The required time to train the various machine learning algorithms using encoding techniques is shown in Fig. 1. Similarly, Fig. 2 also presents the training time of the models using mixed encoding methods. Fig. 1 and Fig. 2 show that the NB model required the lowest amount of time to train among the models using data from all encoding techniques. However, the accuracy of this model is not remarkably lower compared to other models listed in Table 3.

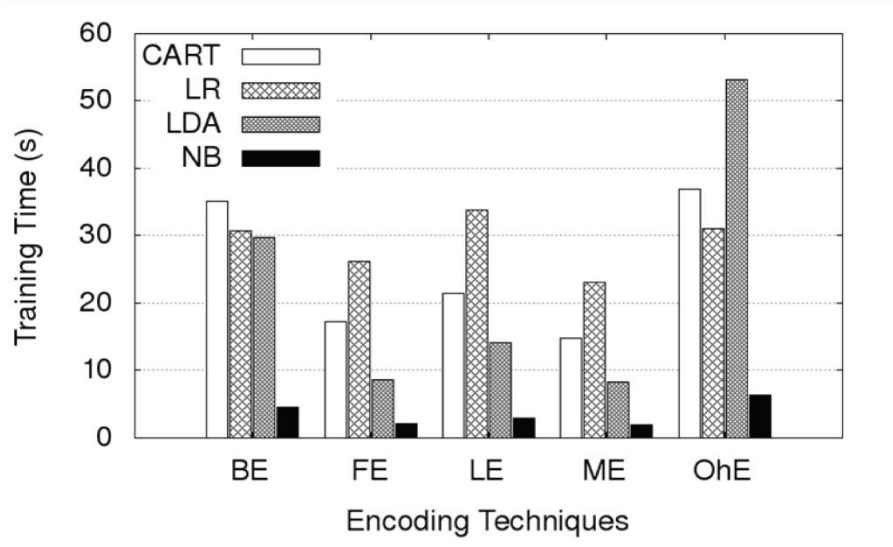

Figure 1. Time Required to Train Various Machine Learning Algorithm Using Encoding Strategies

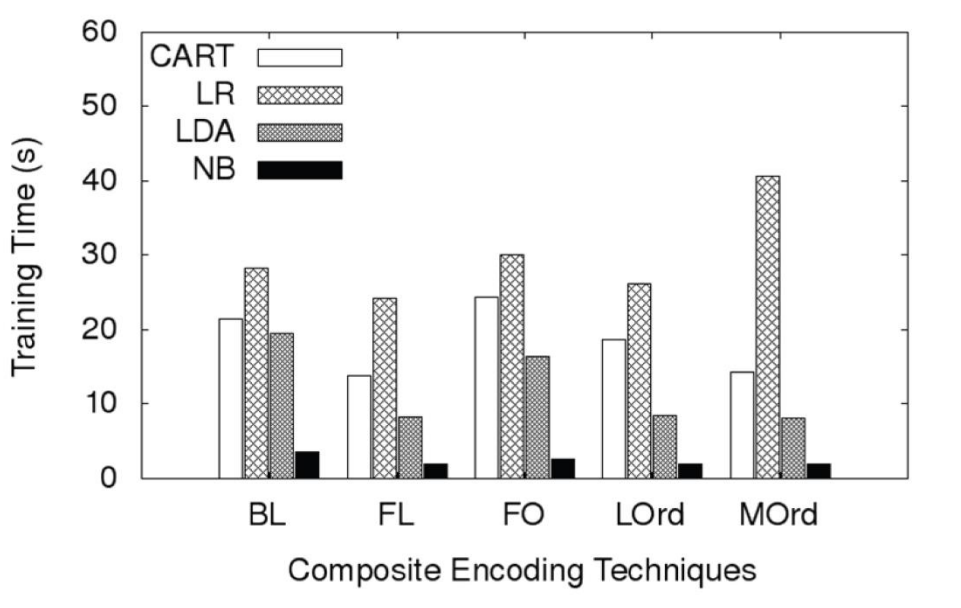

Figure 2. Time Required to Train Various Machine Learning Algorithms Using Composite Encoding Strategies

The Random Forest Classifier took much time to train itself, and its required time is represented in a different Fig. 3. The lowest trained time of RFC was about 500s for Frequency-Label encoding, but that was much higher than other models. Therefore, RFC also would not work better for a large volume of patients' datasets.

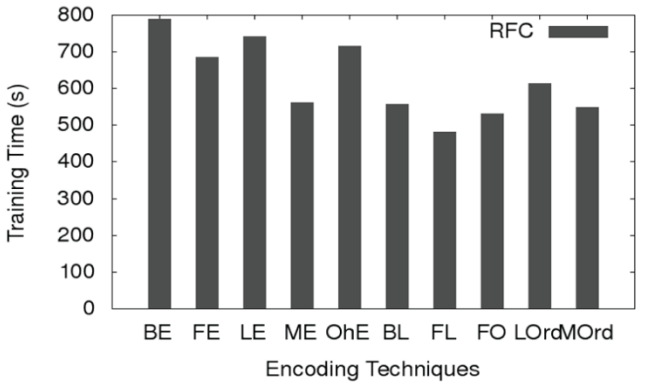

Figure 3. Time Required to Train RFC Using Various Encoding Strategies

The results analysis indicates that KNN and SVM are not suitable for big data. Although the RFC provides better accuracy for a large volume of data, it takes more time to train the machine. By comparing the different models from Table 3, Fig. 1, and Fig. 2, it is found that LDA provides better accuracy (more than 76%) for Binary encoding, Mean encoding, and One-hot encoding with moderate learning time. On the other hand, the Naive Bayes takes the lowest training time for tested datasets, but they achieve less than 74% accuracy for all encoding techniques used here.

6. Conclusion

This article demonstrated and compared the accuracy of various machine learning models applied to categorical features. The categorical variables were encoded using different encoding techniques and the mixed technique of more than one encoding. The goal of this study was to check what learning models comparatively suit better with patients' Big data. Therefore, we find out different metrics, Standard Deviation, and time, required for the machine besides accuracy. The result shows that KNN and SVM are not possible to train against patients' Big data. The RFC is possible to learn, but it takes a long time to be done. Among all models

experimented here, the training time of NB was the lowest for all encoding techniques, but the accuracy was moderated. In all respects, the LDA shows a better performance for the healthcare dataset with average training time. In this study, we also wanted to observe what encoding techniques would help to provide the high accuracy of the trained models. The overall results indicate that the Label encoding technique performs better with lower-dimension.

Acknowledgments

The authors would like to thank the Information & Communication Technology Division, Bangladesh for their financial support.

References

[1] J. Wiens and E. S. Shenoy, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology," *Clin. Infect. Dis.*, vol. 66, no. 1, pp. 149–153, 2018.

[2] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, 2018.

[3] M. D. Ed Corbett, "The Real-World Benefits of Machine Learning in Healthcare.", Health catalyst, 2017.

[4] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, pp. 7–9, 2017.

[5] C. M. Lynch, V. H. van Berkel, and H. B. Frieboes, "Application of unsupervised analysis techniques to lung cancer patient data," *PLoS One*, vol. 12, no. 9, p. e0184370, 2017.

[6] A. So, D. Hooshyar, K. W. Park, and H. S. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Appl. Sci.*, vol. 7, no. 7, p. 651, 2017.

[7] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J. Big Data*, vol. 7, pp. 1–41, 2020.

[8] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv Prepr. arXiv1604.06737*, 2016.

[9] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *J. Big Data*, vol. 6, no. 1, p. 63, 2019.

[10] D. Dua and C. Graff, "UCI machine learning repository." 2017.

[11] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Mach. Learn.*, vol. 107, no. 8–10, pp. 1477–1494, 2018.

[12] R. Karthiga, G. Usha, N. Raju, and K. Narasimhan, "Transfer Learning Based Breast cancer Classification using One-Hot Encoding Technique," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 115–120, 2021.

[13] S. Gupta and R. R. Sedamkar, "Machine Learning for Healthcare: Introduction," in *Machine Learning with Health Care Perspective*, Springer, pp. 1–25, 2020.

[14] S. Raveendran, P. N. Indi, S. Agrahari, S. Menon, and D. A. S. Seelan, "Machine learning based prognostic model and mobile application software platform for predicting infection susceptibility of COVID-19 using health care data," *medRxiv*, 2020.

[15] Z. Sedighi-Maman and A. Mondello, "A two-stage modeling approach for breast cancer survivability prediction," *Int. J. Med. Inform.*, vol. 149, p. 104438, 2021.

[16] P. Mathur, "How to Implement Machine Learning in Healthcare," in *Machine Learning Applications Using Python*, Springer, pp. 37–75, 2019.

[17] C. Wu and C. A. Hargreaves, "Topological Machine Learning for Mixed Numeric and Categorical Data," *arXiv Prepr. arXiv2003.04584*, 2020.

[18] J. M. Magolou-Magolou and A. Ha\"\ir, "Assessment of Lifestyle and Mental Health: Case Study of the FST Beni Mellal," in *International Conference on Business Intelligence*, pp. 83–93, 2021.

[19] J. Johannemann, V. Hadad, S. Athey, and S. Wager, "Sufficient representations for categorical variables," *arXiv Prepr. arXiv1908.09874*, 2019.

[20] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognit.*, p. 107501, 2020.

[21] A. Arora *et al.*, "Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings," *J. Neural Eng.*, vol. 15, no. 6, p. 66028, 2018.

[22] J. T. Hancock and T. M. Khoshgoftaar, "Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection," *SN Comput. Sci.*, vol. 2, no. 4, pp. 1–12, 2021.

[23] A. Ebrahimi, U. K. Wiil, K. Andersen, M. Mansourvar, and A. S. Nielsen, "A Predictive Machine Learning Model to Determine Alcohol Use Disorder," in *2020 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–7, 2020.

[24] M. Abdar, W. Ksi\każek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, p. 104992, 2019.

[25] R. Sujitha and V. Seenivasagam, "Classification of lung cancer stages with machine learning over big data healthcare framework," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 5639–5649, 2021.

[26] J. Wang, F. Deng, F. Zeng, A. J. Shanahan, W. V. Li, and L. Zhang, "Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model," *Am. J. Cancer Res.*, vol. 10, no. 5, p. 1344, 2020.

[27] M. H. Hsieh, L.-M. Sun, C.-L. Lin, M.-J. Hsieh, C.-Y. Hsu, and C.-H. Kao, "Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models," *Cancer Manag. Res.*, vol. 10, p. 6317, 2018.

[28] J. van R. J. V. Geoffrey Holmes Bernhard Pfahringer, "BNG(breast-cancer,nominal,1000000)." .

[29] P. Rodr\'\iguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image Vis. Comput.*, vol. 75, pp. 21–31, 2018.

[30] A. Mahani and A. R. B. Ali, "Classification Problem in Imbalanced Datasets," in *Recent Trends in Computational Intelligence*, IntechOpen, 2019.

[31] R. Dinga, B. W. J. H. Penninx, D. J. Veltman, L. Schmaal, and A. F. Marquand, "Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines," *bioRxiv*, p. 743138, 2019.

[32] A. Tharwat, T. Gaber, A. Ibrahim, A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial", AI communications 30 (2), 169–190, 2017.

[33] Liu, Y. Wang, J. Zhang, "New machine learning algorithm: Random forest", in: International Conference on Information Computing and Applications, Springer, 2012, pp. 246–252.

[34] Y. Zhang, "Support vector machine classification algorithm and its application", in: International conference on information computing and applications, Springer, pp. 179–186, 2012.