An approach of cervical cancer diagnosis using class weighting and oversampling with Keras

Hieu Le Ngoc, Khanh Vo Pham Huyen
Department of Computer Science, Faculty of Information Technology, Ho Chi Minh City Open University,
Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Jun 24, 2021
Revised Nov 06, 2022
Accepted Nov 16, 2022

Keywords:

Cervical cancer diagnosis
Class weighting
Logit model
Machine learning
Oversampling
Prediction of cervical cancer

ABSTRACT

Globally, cervical cancer caused 604,127 new cases and 341,831 deaths in 2020, according to the global cancer observatory. In addition, the number of cervical cancer patients who have no symptoms has grown recently. Therefore, giving patients early notice of the possibility of cervical cancer is a useful task since it would enable them to have a clear understanding of their health state. The use of artificial intelligence (AI), particularly in machine learning, in this work is continually uncovering cervical cancer. With the help of a logit model and a new deep learning technique, we hope to identify cervical cancer using patient-provided data. For better outcomes, we employ Keras deep learning and its technique, which includes class weighting and oversampling. In comparison to the actual diagnostic result, the experimental result with model accuracy is 94.18%, and it also demonstrates a successful logit model cervical cancer prediction.

Corresponding Author:

Hieu Le Ngoc
Department of Computer Science, Faculty of Information Technology, Ho Chi Minh City Open University
Room 605, 35-37 Ho Hao Hon Street, District 1, Ho Chi Minh City, Vietnam
Email: hieu.ln@ou.edu.vn

1. INTRODUCTION

With roughly 10 million deaths each year, cancer is one of the main causes of death in the globe. According to globocan statistics [1], the world has about 1/6 of deaths due to cancer in 2020. The disease manifests itself at a very late stage and the patients who are lacking diagnosis and treatment are very common. The cervix, a woman's opening from the vagina to the uterus, is where cervical cancer grows. It is the 4th disease of cancer based on the number of new cases and the 6th disease of cancer in the number of deaths among women worldwide [2]. Women are most likely to get cervical cancer developing in their 30 s or older, it is the average age of 48–52 years old. Although the disease causes great damage to the uterus, this disease progresses silently for a long time (from 5 to 20 years) and its symptoms are quite faint, easy to get confused with other gynecological diseases. So, it is difficult to detect this kind of disease in the early stages. In the late stages, the patient must undergo a partial or complete removal of the uterus, which directly affects the reproductive organs and motherhood of the woman. At the same time, artificial intelligence (AI) [3] is being applied to diagnose diseases and manage all kinds of health problems. Researching on applying AI, especially machine learning and deep learning techniques in cancer diagnosis, are very popular and developing in recent times.

From these issues above, we realize that building a model to diagnose cervical cancer is a practical work. It can assist people in learning about their risk factors or potential for developing cancer from an early stage. Therefore, they can take appropriate actions and control in time.

Studies on cervical cancer diagnosis through machine learning, most of which focus on the dataset on the causes of cervical cancer risk, were observed at the Universitario de Caracas Hospital in Caracas, Venezuela [4].

When studying this dataset, we found that this data set is imbalanced and has a large bias, so it needs to apply many methods and techniques to process. While Keras [5] is a library written in python language that possesses many good supports for building deep learning models, we considered combining Keras deep learning libraries in handling imbalanced datasets [6], [7]. Based on deep learning we propose using a logit model built from the logistic regression algorithm a binary classification algorithm to diagnose a person with cervical cancer or not, then apply class weighting and oversampling techniques and evaluate its accuracy and performance through evaluation metrics like accuracy, precision, recall, F1 score. With the aid of these methods, it can forecast cervical cancer by learning from the unbalanced data. It is feasible that no studies have yet applied using this kind of method. Therefore, we would like to propose this approach in the detail in the next section.

The recent studies' machine learning algorithms are the premise and foundation for the inspiration from this paper's research direction, such as the research of Ijaz *et al.* [8] in 2020 with a model for predicting cervical cancer using risk factors [4] as inputs. Utilizing noise and isolation forests with density-based spatial clustering of applications, they eliminated outliers, using the synthetic minority over-sampling technique (SMOTE) to balance the data, used the random forest (RF) as the classifier for the model. The model is built with 4 cases that combined the above methods, and finally RF performed the best among all testing models. With the same aims, the next study from Alsmariy *et al.* [9] in 2020 also used [4] to build a categorical model through SMOTE to balance the dataset, principal component analysis to increase model performance and stratified technique k-fold cross-validation to prevent overfitting problems combined with a trio of classifiers: decision tree (DT), logistic regression, and RF. The results show that the high reliability of the model. Alam *et al.* [10] also used SMOTE and a stratified 10-fold cross-validation technique to predict cervical cancer in 2019. The study using three classifications: boosted DT, decision forest, and decision jungle algorithm on [4] which is balanced. The above classifiers with the advantages of requiring fewer data to train as well as fast construction and classification with high accuracy. The paper of Abisoye *et al.* [11] in 2019 proposed a study to predict cervical cancer by combining a genetic algorithm (GA) to select the fittest features for the classification process and a support vector machine (SVM) to build the classifier. The dataset [4] was processed to eliminate noise using the min-max standardized technique, then the suitable features were selected through the GA method. Finally, SVM was used in training the system using prepared data for cervical cancer classification.

Two articles by Nithya and Ilango [12] and Mehmood *et al.* [13] in 2019 build some classifier models based on machine learning algorithms. They use the feature selection technique to select features and the k-fold cross-validation method to prevent overfitting. In the first one article the accuracy of the model increased from 97% to 100%, and in the second get higher metrics of area under the curve (AUC) and F1-score. This paper shows that the better feature, the better model.

Besides cervical cancer, we also consider other cancer that has a similar concept to the article on cervical cancer about how to use the risk factors dataset as input and methods and techniques to build classification models. For example, the paper of Maleki *et al.* [14] in 2020, has applied k-nearest neighbors (kNN) combined with a genetic algorithm to select risk features for lung cancer classification. The model is built with 4 methods: DT, kNN without GA (k = 6, 10), and kNN with GA (k = 6). The result with the accuracy of the classification increased to 100% after applying the above method. Or research by Khalil *et al.* [15] in 2020, is about the development of a fuzzy expert system that is a method for predicting lung cancer with the data is its symptoms. The generalization of cancer prediction paper by Maliha *et al.* [16] in 2019 used the Naïve Bayes, k-NN, and J48 algorithm combined with 10-fold cross-validation to predict 9 types of popular cancers with the dataset containing 61 attributes about symptoms. The Weka tool can then be used to assess the precision of cancer data sets. Finally, a review paper [17] of 30 research papers in 2019, the authors surveyed some past research papers and compared the accuracy of different machine learning (ML) algorithms for cancer prediction depending on given datasets or their properties. Some articles used some classifications such as SVM, RF, Naïve Bayes, DT, KNN, fuzzy neural network.

This paper is organized in 5 sections. The fisrt section talks about the cervical cancer situation, the machine learning techniques implemented in this article, and the most related works on the topic. Section 2 is our proposed work which uses the logit model with the Keras technique to handle imbalanced data, introduce the dataset, and the result of the pre-processing data. The experiment's findings would be presented in section 4. Section 5 concludes by summarizing the study's overall findings.

2. PROPOSED METHOD

2.1. Theoretical basis

Keras is a library written in python, which provides the application programming interface (API) for deep learning. It is built upon the well-known machine learning framework, TensorFlow. Keras helps us in building logit models easily and dealing with imbalanced datasets [18] by using deep learning on its own.

We have numerous options for dealing with imbalanced data in TensorFlow and Keras [7]. But in this paper, we are going to use two main techniques to handle imbalanced data are oversampling and class weights. Also, receiver operating characteristic (ROC) techniques [19] which are used to avoid overfitting in the model are dropout regularization and early stopping.

The metrics are used in this article from a confusion matrix for a binary classifier. The matrix is a table where each column denotes a predicted class and each row denotes an actual class, as shown in Table 1. In addition, we also use accuracy, precision, recall, ROC curve, and precision-recall curve (PRC) [20], [21].

Table 1. The confusion matrix

Predicted class actual class	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

2.2. Model building process

1) Input: patient data
The input of the proposal is the data of a patient. The input data will be handled via the predictive model of the classifier. This model will use the dataset [4] as a training dataset to predict cancer. The essential information of the data is the risk factor features for cervical cancer. These attributes will be detailed described in the next section of this article.
2) Predictive model (building classifier)
The process of building a predictive model of cervical cancer in this proposed method includes 3 steps. The first is to process the data, the second is to build the model, and the last is to predict and evaluate the results. The specific work of each step is described in Figure 1.
   - Step 1: firstly, from the data set of risk factors for cervical cancer [4] we process the data to get a better dataset. In this step, we should clean the data to have better value and it can be fit with the computation. We also need to create training and test sets from the dataset. Finally, the data is normalized for better handling. The step is in detail as: 1) to clean data: the dataset has many missing values, in this case, we used the dataset's statistics as foundation, removed some attributes that were not important or having too many missing values; 2) to split data: we split the dataset into train, test, validation sets; and 3) to normalize data: after splitting the dataset, the training set will be normalized by using the standardscaler [22] technique. By using this method, the mean and standard deviation will both be set to 0 and 1.
   - Step 2: after completing the data processing, we have a complete and normalized dataset. With this new dataset, we will proceed to build a predictive model for the problem of binary classification. We would apply the following two methods to increase the reliability: 1) to apply class weights technique: this will utilize a parameter to pass the Keras weights for each class, this is to make the model "become more attentive" to components with a low positive rate; and 2) to apply oversampling technique: This will expand the training set's sample size by expanding the minority class's sample size.
   - Step 3: we predict the results with the test set and evaluate the accuracy of the model.
3) Output*:* conclude the patient is having cancer or not, with its percentage accuracy.

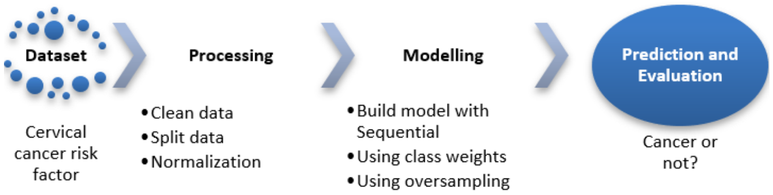


Figure 1. Model building process

2.3. Risk factors cervical cancer dataset

The risk factor cervical cancer dataset used in this article is from the University of California, Irvine (UCI) machine learning repository [23]. The dataset was observed at Hospital Universitario de Caracas in Caracas, Venezuela in 2017. The dataset includes 858 patients' demographic data, lifestyle choices, and previous medical records. Many omitted values exist in this dataset because some patients have decided not to respond a few questions due to their privacy concerns. Table 2 contains the information describing the data type, and how many values are missing for each attribute.

The research’s data collection of cervical cancer risk factors is a small dataset and imbalanced. The attribute diagnosed (Dx): cancer is used to classify cancer with the number of patients diagnosed with cancer being very small, a statistic is presented in Figure 2. With a sample of 858 patients, only 18 people were diagnosed with cancer, respectively 2.1% of the total number of patients. This highly imbalanced effect will affect the model’s predicted quality. In addition, Table 1 also shows that the quantity of missing values in the dataset is very high (up to 26 out of 36 missing value properties) with the largest 2 missing attributes being 787 and 18 missing attributes over 100 values.

Table 2. Attributes of the cervical cancer dataset

No.	Attribute	Type	No. of missing value	No.	Attribute	Type	No. of missing value
1	Age	Integer	0	19	STDs: pelvic inflammatory disease	Bool	105
2	Number of sexual partners	Integer	26	20	STDs: genital herpes	Bool	105
3	First sexual intercourse	Integer	7	21	STDs: molluscum contagiosum	Bool	105
4	Num of pregnancies	Integer	56	22	STDs: acquired immunodeficiency syndrome (AIDS)	Bool	105
5	Smokes	Bool	13	23	STDs: human immunodeficiency virus (HIV)	Bool	105
6	Smokes (years)	Float	13	24	STDs: hepatitis B	Bool	105
7	Smokes (packs/year)	Integer	13	25	STDs: human papillomavirus (HPV)	Bool	105
8	Hormonal contraceptives	Bool	108	26	STDs: number of diagnoses	Integer	0
9	Hormonal contraceptives (years)	Float	108	27	STDs: time since first diagnosis	Integer	787
10	IUD (intrauterine device)	Bool	117	28	STDs: time since last diagnosis	Integer	787
11	IUD (years of intrauterine device)	Float	117	29	Dx: cancer	Bool	0
12	Sexually transmitted diseases (STDs)	Bool	105	30	Dx: cervical intra-epithelial (CIN)	Bool	0
13	STDs (number)	Integer	105	31	Dx: HPV	Bool	0
14	STDs: condylomatosis	Bool	105	32	Dx	Bool	0
15	STDs: cervical condylomatosis	Bool	105	33	Hinselmann	Bool	0
16	STDs: vaginal condylomatosis	Bool	105	34	Schiller	Bool	0
17	STDs: vulvo-perineal condylomatosis	Bool	105	35	Cytology	Bool	0
18	STDs: syphilis	Bool	105	36	Biopsy	Bool	0

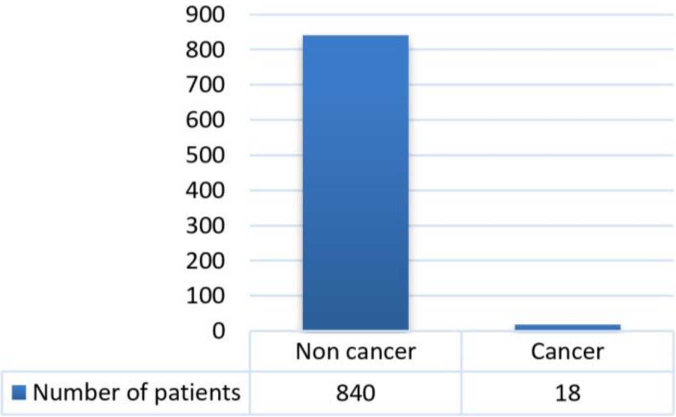


Figure 2. The small fraction of positive samples (cancer diagnosed)

2.4. Data processing

1) Clean data

- Remove the 6 diagnostic attributes: Dx: CIN, Dx: HPV, Dx: hinselmann, schiller, cytology, biopsy. However, we keep the Dx: cancer attribute as the label for classify process.
- Remove the 2 attributes with many missing values: STDs: time since first diagnosis and STDs: time since last diagnosis.
- Missing value*:* for attributes of true/false meaning, we replace missing values with a value of “0”. For attributes that have a quantity meaning (numeric type), We substitute missing values with the characteristics’ average values, where appropriate. After cleaning, the new dataset has 27 attributes, 26 attributes are risk factors, and 1 attribute is for classifying. All the attributes do not contain any missing value.

2) Split data
Three subsets of the dataset were created: training set, validation set, and testing set. Specifically, with the sample of 858 tuples, the training set and testing set are separated from the dataset. The testing set is 20% of the dataset. The training set makes up the remainder of the dataset, which takes 80%. This is a large training set, and it will be divided into 2 more subsets, the validation set and the smaller training set. The validation set also is 20% of this large training set. Each subset is divided into labels set (containing the model's classification class, the Dx: cancer attribute) and the features set (containing 26 attributes about cervical cancer risk factors). As a result, we get the shape of the labels and features of the training, validation, and test set respectively as (548), (138), (172), (548, 26), (138, 26), (172, 26).

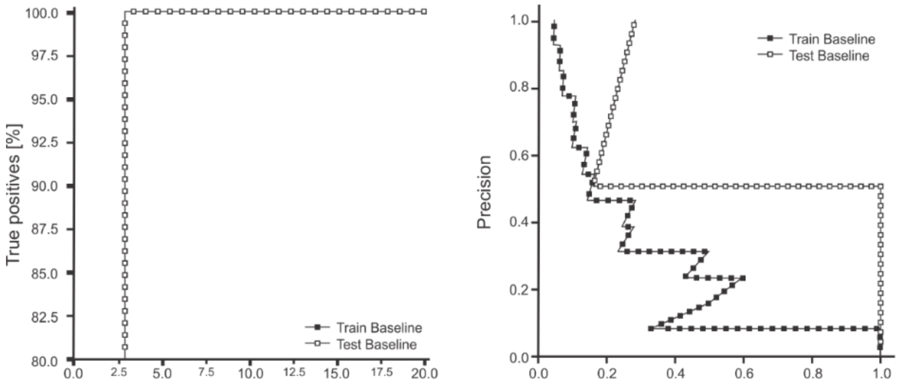
3) Normalize data
We use StandardScaler to normalize the input features. The standard deviation will be set to 1 and the mean to 0. This technique is provided by Keras and it is simply to use.

3. RESULTS AND DISCUSSION

3.1. Baseline model

Create a baseline model [24] by sequential [25] with 3 layers: dense, dropout (hidden layer), dense 1 (as output). The hidden layer relates to an input layer, a dropout layer is used to lessen overfitting, and an output sigmoid layer is the ratio that the patient gets cancer. The model uses epochs = 40, batch size = 50 and total params is 449. In addition, we also set the bias to reduce the loss of the model. We used Keras to evaluate the baseline model with the testing set, with a sample of 172 patients, the model correctly predicted 170 non-cancer patients (true negatives) and failed to predict 2 cancer patients (false negatives), 98.8% in accuracy. In this case, the results show that there are relatively few false predictions, which means that there are relatively few cases of the disease that go undetected. However, we may want to have fewer false negatives despite the number of false positives is rising. This compromise may be more appropriate since it is safer to have fewer false negatives than to have more false positives. The results are shown by the ROC and area under precision recall curve (AUPRC) charts [26].

The Figure 3(a) shows a low false-positive rate of around 2.5%. It is known that the testing set has errors, but the training set does not. In Figure 3(b), the precision and recall distributions of the baseline model are shown. For the training set, if the precision decreases, the recall will increase, the curve of testing baseline set is a convex and concave polygon line due to the imbalanced dataset. This shows that the dataset is not good, and we need to use some other techniques to ameliorate the fitness of the model.



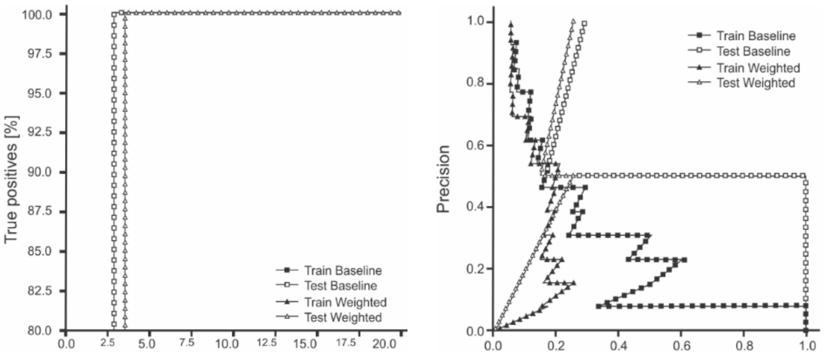
(a) (b)

Figure 3. The baseline model results with (a) ROC chart and (b) AUPRC chart

3.2. Model improvement with class weighting

To improve the fit of the model, we use the class weighting technique provided by Keras. After setup and calculation according to the guide of Keras, the weights of the model with classes 0 and 1 are 0.51 and 23.83, respectively. With the sample of 172 tuples, the model has predicted 160 non-cancer patients (true negatives), 10 cancer patients but incorrectly (false positives) and 2 patient cancers (true positives) with an accuracy is 94.18%. Although this model incorrectly predicted up to 10 people who did not get cancer, in exchange, it can find people with the disease. In this case, we determine whether the patient has this disease is much more important than misdiagnosing.

After using class weighting, the model result is quite smooth and fits better with the dataset. Take a look at Figure 4, Figure 4(a) shows that the ratio of false positive after using the class weighting of the testing set has increased and was greater than 2.5% and approximately 3.0%. The Figure 4(b), it is so obvious that the precision and recall distributions are more reasonable, the testing set after weighting is closer to the training set, showing an increased model fitness.



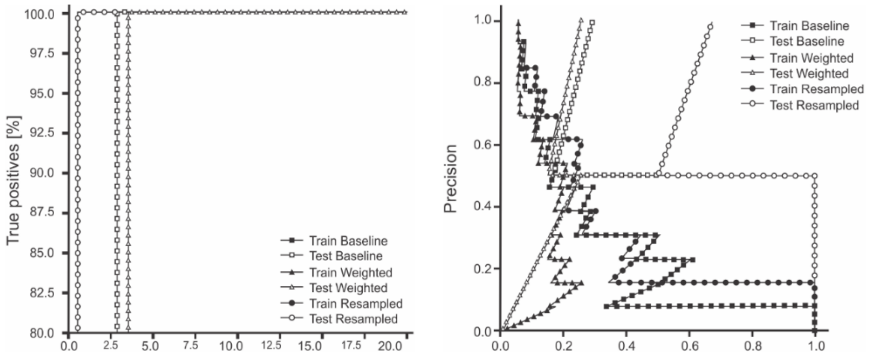
(a) (b)

Figure 4. Comparing the baseline model result and weighted model result in: (a) ROC chart and (b) AUPRC chart

3.3. Model improvement with oversampling

In the training set, the quantity of positive tuples is 15 and the number of negative samples is 535. The oversampling method has increased the number of positive samples of the training set so that it is equal to the number of negative samples. Then, we will combine this new positive sample with the negative sample, so we have a new training set that is increased in size to 1070 records. This oversampling model predicted 134 non-cancer patients, 36 cancer patients but incorrectly, and correctly predicted 2 cancer patients with an accuracy of 79.07%. The results of the model are not very satisfactory compared to the class weighting method.

After applying the oversampling technique, the resampled model also seems smoother and more fit with the data, compared to the baseline model. In the left chart of Figure 5, the false positive ratio of the resampled model is the smallest among the 3 models. With the AUC point is the greatest among them, the resampled model shows us the most trusted model. But in the right chart of Figure 5, the distribution curve of precision and recall of the 3 models, shows us a clearer picture of these techniques. The lines are for the resampled model, we can see these 2 lines seem to be parallel and this represents the fittest model. But the distance between these 2 lines is the longest one among the 3 of them. With the application of the 2 techniques, we can see they are better than the original one, but each of these 2 techniques also has its own pros and cons. So far, in this experiment result, we can see what we should do when we use the class weighting or oversampling technique.



(a) (b)

Figure 5. Comparing the baseline model result, weighted model result and resampled model result in:
(a) ROC chart and (b) AUPRC chart

When applying the oversampling technique, the resampled model also seems smoother and more fit with the data, compared to the baseline model. In Figure 5(a), the false positive ratio of the resampled model is the smallest among the 3 models. With the AUC point is the greatest among them, the resampled model shows us the most trusted model. But in Figure 5(b), the distribution of precision and recall curve of the 3 models, shows us a clearer picture of these techniques. The lines are for the resampled model, we can see these 2 lines seem to be parallel and this represents the fittest model. But the distance between these 2 lines is the longest one among the 3 of them. With the application of the 2 techniques, we can see they are better than the original one, but each of these 2 techniques also has its own pros and cons. In brief, in this experiment result, we can see what we should do when we use the class weighting or oversampling technique.

4. CONCLUSION

This article is a research result about the logit model, imbalanced data, and some deep learning techniques is provided by Keras. We proposed a cancer disease prediction system that is used the model built by the imbalanced dataset which contains the risk factors of cervical cancer. With the 3 models with different techniques (baseline, class weighting and oversampling), the program runs out the diagnostic result coincides with the fact that 2 records correctly reached the highest rate of 94.18% with the weighted model. This initial information can be a warning which will be the basis for the patient to have an objective view of their disease, so they can make plans to care for and to take treatment that is appropriate. For further development, we aim to collect the most suitable dataset or process the dataset fitter for better prediction. Beyond that, we also need to study more about other deep learning techniques in imbalanced dataset, and some improved logit models.

ACKNOWLEDGEMENTS

We would like to express our big appreciation to Ho Chi Minh City Open University in Vietnam for supporting us in our research on this topic.

REFERENCES

[1] "Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey," *World Health Organization*, 2020. https://apps.who.int/iris/handle/10665/331452 (accessed Mar. 15, 2021).

[2] "Cancer fact sheets," *Cancer today; World Health Organization*, 2021. https://gco.iarc.fr/today/fact-sheets-cancers (accessed Mar. 15, 2021).

[3] B. J. Copeland, "Artificial intelligence," 2022. https://www.britannica.com/technology/artificial-intelligence (accessed Jul. 13, 2022).

[4] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," 2017, pp. 243–250, doi: 10.1007/978-3-319-58838-4_27.

[5] F. Chollet, "Keras." 2015. https://github.com/fchollet/Keras (accessed Apr. 16, 2022).

[6] "Classification on Imbalanced data," *Tensorflow*. https://www.tensorflow.org/tutorials/structured_data/imbalanced_data (accessed Apr. 07, 2021).

[7] "10 techniques to deal with Imbalanced classes in machine learning." Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/ (accessed Nov. 03, 2022).

[8] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, 2020, doi: 10.3390/s20102809.

[9] R. Alsmariy, G. Healy, and H. Abdelhafez, "Predicting cervical cancer using machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020, doi: 10.14569/IJACSA.2020.0110723.

[10] T. M. Alam, M. M. A. Khan, M. A. Iqbal, A. Wahab, and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019, doi: 10.14569/IJACSA.2019.0100251.

[11] Abisoye O. A., Abisoye B. O., E. Ayobami, and K. Lawal, "Prediction of cervical cancer occurrence using genetic algorithm and support vector machine," *3rd International Engineering Conference (IEC 2019)*, 2019. [Online]. Available: http://repository.futminna.edu.ng:8080/jspui/bitstream/123456789/11397/1/PredictionofCervicalCancer.pdf

[12] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Applied Sciences*, vol. 1, 2019, doi: 10.1007/s42452-019-0645-7.

[13] M. Mehmood, M. Rizwan, M. Gregus ml, and S. Abbas, "Machine learning assisted cervical cancer detection," *Frontiers in Public Health*, vol. 9, 2021, doi: 10.3389/fpubh.2021.788376.

[14] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Systems with Applications*, vol. 164, 2021, doi: 10.1016/j.eswa.2020.113981.

[15] A. M. Khalil, S. -G. Li, Y. Lin, H. -X. Li, and S. -G. Ma, "A new expert system in prediction of lung cancer disease based on fuzzy soft sets," *Soft Computing*, vol. 24, pp. 14179–14207, 2020, doi: 10.1007/s00500-020-04787-x.

[16] S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollick, and T. Islam, "Cancer disease prediction using naive bayes, K-Nearest neighbor and J48 algorithm," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–7, doi: 10.1109/ICCCNT45670.2019.8944686.

[17] A. Kumar, R. Sushil, and A. K. Tiwari, "Machine learning based approaches for cancer prediction: A survey," *2nd International Conference on Advanced Computing and Software Engineering (ICACSE-2019)*, 2019, doi: 10.2139/ssrn.3350294.

[18] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[19] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv*, 2017. [Online]. Available: https://arxiv.org/pdf/1710.10686.pdf

[20] Hossin M. and Sulaiman M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 1–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[21] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.

[22] "Sklearn.preprocessing.standardscaler," *scikit-learn*, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed May 15, 2020).

[23] "UCI machine learning repository," Irvine, CA: University of California School of Information and Computer Science, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/

[24] "Baseline model," *Science direct*. https://www.sciencedirect.com/topics/computer-science/baseline-model (accessed Jun. 07, 2021).

[25] Fchollet, "The sequential model," *Keras*. https://Keras.io/guides/sequential_model/ (accessed Apr. 20, 2020).

[26] W. J. Krzanowski and D. J. Hand, *ROC curves for continuous data*, New York: Taylor & Francis, 2009, doi: 10.1201/9781439800225

BIOGRAPHIES OF AUTHORS

Hieu Le Ngoc has been working in IT industry as an IT System Architect since 2010. As now, he is working as an IT instructor for HCMC Open University (Vietnam). His major study is about cloud computing, cloud efficiency for better service, AI's application, Machine Learning and Computer Linguistics. His minor studies are education, education in IT line, Languages Teaching (Chinese & English), Business and Economic. He can be contacted at email: hieu.ln@ou.edu.vn and ResearchGate: https://www.researchgate.net/profile/Hieu-Le-24.

Khanh Vo Pham Huyen is an IT student at Ho Chi Minh City Open University. She is getting Bachelor of IT in September 2021. She has been starting to study about data and machine learning from 2020. She is keen on medicine data and how to discover a disease with data. She would go on with further studies and keep moving up to post-graduation. She can be contacted at email: 1751010056khanh@ou.edu.vn and ResearchGate: https://www.researchgate.net/profile/Khanh-Vo-Pham-Huyen.