ISSN 2347 - 3983

Volume 9. No. 8, August 2021

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter02982021.pdf

https://doi.org/10.30534/ijeter/2021/02982021

A Survey on Data Mining Techniques for COVID Prediction

Jayshree Pawar[1], Urjita Thakar[2]

[1] Research Scholar, Department of Computer Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, India, jayhreepawar.1010@gmail.com

[2] Professor, Department of Computer Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, India, urjita@rediffmail.com

ABSTRACT

Corona Virus Disease of 2019 (COVID-19) has emerged as a serious health emergency worldwide. The symptoms of COVID-19 are un-detectable at early stage in most of the patients. It spreads from person to person very rapidly and causes severe sickness and loss of life in a number of cases if not treated early. Data mining techniques are very commonly being used in medical sector for detection and prediction of a variety of diseases and medical conditions of patients. A number of researchers are also working towards prediction of possibility of infection of COVID-19 among humans using machine learning techniques, specifically by applying data mining methods. In this paper, an extensive survey of available literature in the domain of prediction of COVID-19 infection and other diseases has been presented. This also includes survey on data mining techniques, models and various datasets.

Key words : Data Mining, Machine learning, COVID-19, Prediction, Diagnosis, Feature Selection, Misclassification.

1. INTRODUCTION

Coronavirus epidemic has grappled the whole world. Countries on all the continents are fighting to save their citizens from this deadly disease. The World Health Organization revealed the official name of the pneumonia transmitted by this virus as "COVID-19" or "Corona Virus Disease 2019" on February 11, 2020 [1]. Corona virus is an infection transmitted by a novel severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2). The virus spreads rapidly among people in many different ways which is a major concern around the world [2]. Though it spreads primarily through the air [3]. According to a report, World Health Organisation first identified this virus on 31 Dec 2019 in Wuhan city, China. Many of the first cases of COVID-19 were related to Huanan seafood wholesale market, implying that SARS-CoV-2 was spread from animals to humans [4]. This pandemic is a major public health issue that is affecting people all over the world. Also, it is a contagious disease and causes severe sickness and loss of life in a number of cases if not treated in early stages. According to a study in which parametric analysis was carried out, COVID-19 has a growth rate that is roughly twice that of SARS and MERS [5].

Health care workers are working over time for more than past one year to fight against this deadly virus. Every day, the health-care industry is generating massive amount of data about COVID patients and disease. Researchers and physicians across the globe are working together to detect the infection in people at early stages and find a treatment to cure this disease.

Some researchers have analysed severity of COVID infection based on specific existing illness condition such as cancer, pneumonia, pregnancy, hypertension etc. Also different types of datasets have been used by different researchers. These include images from chest X-ray, CT-Scans, pathological reports of patients etc. Some researchers are focusing on test methods and attempting to minimise testing workload [6].

Many researchers are using machine learning methods, especially data mining techniques in the healthcare domain. These are used to discover useful information out of a huge amount of data and to present it in an easy-to-understand format for humans. Classification and clustering are among the most common data mining techniques. Disease prediction is very significant application of these techniques.

The algorithms of machine learning are very important for the diagnosis of diseases and they have a significant impact in the medical field. Medical data mining is a term used to describe a variety of strategies for discovering valuable patterns that assist in medical diagnosis. This is aimed at improved disease prediction and early diagnosis. This aids in faster and better medical treatment and patient care [16].

Despite the fact that several studies have been conducted on prediction of COVID and other diseases using various machine learning techniques and variety of datasets, very little literature containing a survey of these is available. In this work, therefore, a review of numerous studies conducted on the prediction of COVID and other diseases is presented.

Rest of the paper is organized as follows: In section II, recent study on COVID prediction is given. In section III, the work done to predict various diseases is discussed. In section IV, feature selection techniques presented by different researchers is discussed. In section V, the literature on data imbalancing techniques is discussed. Finally, in section VI the conclusion and discussion is presented.

2. RECENT STUDY ON COVID PREDICTION

In past one and half years lot of literature has been generated in the field of recognition and prediction of COVID disease by using number of machine learning techniques and different datasets.

At the early time of spread of COVID in the year 2020, Naoya Itoh et al. studied the COVID immune responses, genomics, identification, treatment and management of the disease. They also reviewed the prevention and control strategies for such disease. They advised that globally countries need to pay more attention to corona virus disease monitoring systems and increase country's readiness [2].

To detect coronavirus disease, CT scans, Magnetic Resonance Imaging (MRI), and other imaging techniques are beneficial. Medical images assist physicians in determining the impact of disease on affected persons. This type of data is extremely useful to detect the accuracy and efficiency of diagnosis [6]. Therefore, large amount of data containing medical images are available.

COVID prediction from X-rays using Artificial Intelligence (AI) techniques can be extremely useful, and it could be able to alleviate the shortage of doctors and physicians in rural areas. F. Pan et al. examined the improvements in the lungs of COVID patients from initial stage of diagnosis to recovery of patient using the chest CT findings [7]. Radiologists differentiated COVID from other viral pneumonia using chest CT scans with a high degree of specificity [8]. D. Haritha et al. presented a transfer learning method for predicting coronavirus cases from images of patient's chest X-ray [9]. In another work, authors used Computed Tomography (CT) images. They constructed an ensemble model using multivariate logistic regression by combining the features of radiomics and deep learning to distinguish critical cases from severe cases of COVID [10].

With chest X-rays, S. Rajaraman et al. showed COVID-19 pulmonary manifestation detection using iteratively pruned deep learning model ensembles. To minimise complexity and increase memory performance, the best performing classifiers were pruned iteratively. To boost classification accuracy, authors performed predictions by combining the pruned models using various ensemble strategies. Improved predictions were achieved by combining modality-specific information transfer, iterative process pruning, and ensemble learning [11].

R. Kumari et al. examined some established forecasting models in depth and predicted the number of confirmed, recovered, and death cases caused by COVID-19 in India. This research looked into how COVID-19 spread through India. The possibility of cases that may arise in the future was predicted using multiple linear regression and autoregression. This prediction may be useful in resource control, such as health care and prompt steps can be taken with advance planning to minimise human life loss [6].

In a report, Epidemiologists predicted COVID-19 confirmed cases which could rise in the United States and various other nations. Authors used two unsupervised classic clustering approaches: K-means clustering and correlation to forecast the distribution of disease. They also predicted a 0.85 relationship between overall deaths and critical patient attributes [12].

V. Bhadana et al. compared five machine learning standard models to forecast the threatening variables of COVID-19: linear regression (LR), decision tree, least absolute shrinkage and selector operator, random forest, and SVM. Each model generated three types of forecasts for COVID prediction in the next five days: total active cases, total deaths, and total recoveries. Authors analysed the findings of the experiment in which poly LR, LASSO, Random forest, and decision tree showed better results and SVM showed a weak outcome [13].

E.V. Robilotti et al. analysed the risk factors on COVID-19 for severe infection in patients with more than one illness including cancer. Authors proposed that the result of coronavirus was more severe among people with cancer [14].

Early risk identification of COVID can be done byexamining three primary sounds: coughing, breathing, and voice [1]. A number of researchers have analysed the features of cough, breathing, and voice of the patients using the Recurrent Neural Network (RNN), Convolution Neural Network (CNN), Artificial Neural Network (ANN) and specifically its important well-known architecture, the Long-Short Term Memory (LSTM) [15], [16], [17]. As opposed to coughing and breathing sound samples, the speech test had a poor accuracy [15].

Based on the recent researches it has been observed that various forecasting models have been developed to predict COVID. Authors used chest X-ray images, cancer reports of patients to predict the possibility of infection. The authors demonstrated the ability of machine learning models to predict the number of future covid-19 infected patients, which is now considered as a significant challenge to humanity.

3. DISEASE PREDICTION USING DATA MINING

Earlier researchers have done lot of work in the area of disease prediction. Contribution of various researchers is discussed next.

Disease Prediction is a popular application of data mining. Various types of diseases such as liver disorder, diabetes, breast cancer, thyroid illness, skin cancer, etc. can be predicted using data mining.

R. Vijiyarani et al. presented various algorithms of machine learning used in the area of disease prediction. The focus of survey presented was use of data mining techniques and multiple target attributes to predict various types of diseases namely Heart disease, Diabetes and Breast cancer disease predictions [18]. Some authors focused on the decision parameters, attributes and features that are used to predict disease. They highlighted the usefulness of various classification systems for disease detection in medical datasets [19].

A number of researchers investigated and compared different data mining and machine learning techniques such as hybrid ANN, back propagation, Decision Tree, Random Forest, Naive Bayes, KNN, association rule etc. for prediction of heart disease [20]-[23].

Diabetes affects millions of people around the world. Many of these individuals are completely unaware of their condition. Machine learning models such as Logistic Regression, AdaBoost, etc. for prediction of diabetes have also been proposed by earlier researchers. The authors also highlighted the significance of various classification methods used for disease prediction in medical datasets [24]. N. Nnamoko et al. researched to predict diabetes and exploit diversity from heterogeneous base classifiers and the optimisation effect of attribute subset selection in order to improve accuracy [25]. In another work, authors predicted whether someone has diabetes or not. Authors applied a new algorithm called the Homogeneity-Based Algorithm (HBA) and combined it with existing classification algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), or Decision Tree (DT) to improve classification accuracy. The outcomes of experiment showed that the suggested approach beats current approaches significantly [26].

R. Baitharu et al. examined the effect of liver disorder using six different classifiers namely J48, Naive Bayes, Multilayer Perceptron, IBK, ZeroR, VFI. The findings indicated that all of the classifiers had improved predictive performance, however Naive Bayes showed the weak outcome [27].

Breast cancer has surpassed all other cancers as the leading cause of mortality among women. Shelly Gupta et al. studied the literature available on diagnosis and prognosis of breast cancer using machine learning models such as Multilayer Perceptron, Self Organizing Map, Radial Basis Function, Artificial Neural Network, Support vector machine etc. Authors analyzed that ANN gave better accuracy as compare to other classification techniques [28].

Authors Presented comprehensive literature review of various deep learning algorithms for early risk identification and classification of skin cancer. Researchers used different algorithms such as Artificial Neural Network, Convolutional Neural Network, Kohonen Self-Organizing Neural Network, and Generative Adversarial Neural Network for classification of lesion images. Studies revealed that CNN performed better as compared to other algorithms for classification of image data [29].

The analysis of the literature reveals that researchers have used several machine learning methods for diagnosing various types of diseases. The classification model's accuracy ranged between 85% to 98%.

In next section, the work carried out for feature selection by different researchers is presented.

4. FEATURE SELECTION FOR CLASSIFICATION

In medical datasets, sometimes data is available in raw form and it may also contain irrelevant features. It is rare that all the features in the dataset are useful to build a machine learning prediction model. Therefore, dimensionality reduction or feature selection becomes an important step before applying any data mining technique to decrease the model's overall complexity. Many researchers have applied feature selection techniques to identify relevant features to improve the quality of dataset. Their contribution is discussed further.

A number of methods are available for feature selection in which dimensionality reduction is the one which is most widely used method for removing noisy and redundant elements. Feature extraction and feature selection are the two major types of dimensionality reduction techniques. Feature extraction usually transforms or combines the original feature space to obtain a new feature space with lower dimensions. While feature selection selects a subset of features from original feature space without performing transformation. Both the methods have the potential to improve learning efficiency, reduce computational complexity, create more generalizable models, and reduce storage requirements [30]. Some authors discussed and compared existing feature selection algorithms such as filter, wrapper and embedded methods with their pros and cons. Their findings indicated that embedded method performed better [31].

To achieve substantial dimensionality reduction in medical datasets, the authors presented an embedded hybrid feature selection model which combines two well known data mining techniques which are clustering and classification. Authors conducted an experiment using F-score and k-means feature selection techniques with SVM classifier on Diabetes, Breast Cancer and Heart Disease datasets. Their findings indicated that choosing the most important features from medical data improve the accuracy of classifier and also helps the physician to make accurate diagnosis [32].

IhsanAbodKhalaf et al. proposed the network intrusion detection system which is based on Support Vector Machines (SVM) classifiers and two feature selection algorithms:- Self Organizing Map (SOM) and Principle Component Analysis (PCA). Voting technique was used to combine the two feature selection algorithms. The findings of this study revealed that different feature selection algorithms can have different impact on classification performance. In addition, authors presented a comparative study of accuracy results between the feature selection algorithms namely Principal Component Analysis and Self Organizing Map using SVM classifier [33].

A. Jovic et al. summarized the application domains of multi-dimensional feature space such as text mining, image processing and computer vision, bioinformatics, industrial applications. Detailed study of feature selection techniques was also presented [34].

In the surveyed literature, feature selection algorithms such as filter, wrapper and embedded methods along with their pros and cons are discussed. Many methods are found to be effective for feature selection in data related to COVID-19.

5. DATA IMBALANCING

The available datasets on COVID are sometimes highly imbalanced, using such datasets can be very difficult. There are some standard techniques to balance the dataset, which are discussed in this section.

Misclassification problem or class imbalance occurs when one class has fewer training instances than the other classes. Phung et al. looked at a variety of approaches and strategies to deal with the issue of class imbalance at both data level and algorithmic level. Authors argued that sampling is one of the most popular method to address the issue. They described fundamental sampling techniques such as undersampling and oversampling, as well as advanced sampling techniques for minimising misclassification problem in training data with their pros and cons. Also, they presented a new technique to handle with the issue of misclassification by integrating supervised and unsupervised learning. These approaches change the majority and minority class distributions in the training data sets to achieve equal number of instances in each class [35].

G.M. Weiss et al. analysed how data imbalancing affects classifier learning. Authors discussed how misclassification influenced learning and how it impacted the evaluation of learned classifiers. The results of two experimental studies are then presented. In the first experiment, authors compared the performance of classifiers built from unbalanced datasets against classifiers built from balanced copies of the same datasets. Their study suggested that original class distribution is frequently unsuitable for learning and better performance can be achieved by using class distribution techniques. In the second experiment, authors suggested which distribution is optimal for training based on two performance metrics: classification accuracy and area under the ROC curve [36]. Many researchers have described most well-known data reduction techniques [37], [38]. J. Laurikkala studied three methods namely Simple random sampling (SRS), One-Sided Selection (OSS), Neighborhood Cleaning rule (NCL) for improving identification of misclassification problem. Authors conducted an experiment with ten datasets, six of which were medical data, which is our major application area of concern. Their experiments indicated that Neighborhood Cleaning rule (NCL) showed better results as compare to simple random sampling and one-sided selection methods. The findings suggested that NCL can be used to improve the modelling of problematic small classes and to create classifiers that can detect these classes from real-world data [39].

Another approach used an over-sampling approach SMOTE to handle the issue of data imbalance. Some authors demonstrated that combination of over-sampling the minority class with under-sampling the majority class will improve classifier efficiency over simply under-sampling the majority class [40], [41].

Studies revealed that real world datasets often have misclassification problem. It occurs when the distribution of classes is biased in the training dataset. Data balancing techniques discussed in the literature could be useful to handle misclassification issue in data available on COVID.

6. CONCLUSION AND DISCUSSION

Corona virus is one of the major cause of death around the world. It's early detection is essential for better medical treatment. In this paper, the available literature towards various diseases, specifically COVID-19 has been reviewed. According to the findings, data mining plays a significant role in disease prediction. Using machine learning to analyse the prediction model yields promising results with better accuracy. Mining the necessary knowledge from medical data assists in making possibly the best diagnosis and decisions. Also, different feature selection and data imbalancing techniques have been presented in a number of papers surveyed. Based on the findings, it has been observed that by using balanced data and reducing the number of attributes, the classification accuracy can be increased. This survey will guide in determining the best algorithm for COVID prediction in order to improve classification. Since the corona virus is still prevailing in the world, mutating its form and causing havoc, it is expected that more machine learning and deep learning methods will be applied to make long term and short term prediction of infection in human population.

REFERENCES

1. [Online] WHO Health information and resources https://www.who.int/emergencies/diseases/novelc-orona virus-2019 [accessed on 1 april 2021].
2. Naoya Itoh et al., "**Coronavirus disease 2019 (COVID-19): A literature review**", Journal of Infection and Public Health, Volume 13, Issue 5, ISSN 1876-0341, 2020.
3. Morawska, Lidia, and Junji Cao. "**Airborne transmission of SARS-CoV-2: The world should face the reality**", Environment international, volume 139 : pages 105730, ISSN: 0160-4120, 2020.
4. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. "**Early transmission dynam-ics in Wuhan, China, of novel coronavirus-infected pneumonia**", N Engl J Med; Volume 382(13):1199207, 2020.
5. Liang K. "**Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS**", Infection, Genetics and Evolution; Volume 82:104306, PMID: 32278147; PMCID: PMC7141629, 2020.
6. R. Kumari et al., "**Analysis and predictions of spread, recovery, and death caused by COVID-19 in India**", Big Data Mining and Analytics, vol. 4, no. 2, pp. 65-75, June 2021.
7. F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang, et aI., "**Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COYID-19) pneumonia**", Radiology, vol. 295, no.3, pp.715-721, 2020.
8. X. Bai et al., "**Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT**", Radiology, pp. 200823, 2020.
9. D. Haritha, N. Swaroop and M. Mounika, "**Prediction of COVID-19 Cases Using CNN with X-rays**", 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, pp. 1-6, 2020.
10. C. Li et al., "**Classification of Severe and Critical Covid-19 Using Deep Learning and Radiomics**", IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3585-3594, Dec. 2020.
11. S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio and S. K. Antani, "**Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays**", IEEE Access, vol. 8, pp. 115041-115050, 2020.
12. R. Kurniawan, S. N. H. Sheikh Abdullah, F. Lestari, M. Z. A. Nazri, A. Mujahidin and N. Adnan, "**Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries**", 8th International Conference on Cyber and IT Service Management (CITSM), Pangkal, Indonesia, 2020, pp. 1-5, 2020.
13. V. Bhadana, A. S. Jalal and P. Pathak, "**A Comparative Study of Machine Learning Models for COVID-19 prediction in India**", IEEE 4th Conference on Information & Communication Technology (CICT), Chennai, India, pp. 1-7, 2020.
14. Elizabeth V. Robilotti, N. Esther Babady, Peter A. Mead, Thierry Rolling, "**Determinants of COVID-19 disease severity in patients with cancer**", Nature Medicine, volume 26, 2020.
15. A. Hassan, I. Shahin and M. B. Alsabek, "**COVID-19 Detection System using Recurrent Neural Networks**" 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), Sharjah, United Arab Emirates, pp. 1-5, 2020.
16. G. Deshpande and B. Schuller, "**An Overview on Audio, Signal, Speech, & Language Processing for COVID-19**", ArXiv, pp. 1-5, May 2020.
17. C. Bales et al., "**Can Machine Learning Be Used to Recognize and Diagnose Coughs?**", ArXiv, pp. 1-10, May 2020.
18. S.Vijiyarani, S. Sudha, "**Disease Prediction in Data Mining Technique – A Survey**", International Journal of Computer Applications & Information Technology Vol. II, Issue I, (ISSN: 2278-7720), January 2013.
19. A. Tikotikar and M. Kodabagi, "**A survey on technique for prediction of disease in medical data**", International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, pp. 550-555, 2017.
20. Syed.Matheen. Pasha, Shilpa Ankalaki, " **Diabetes and Heart Disease Prediction Using Machine Learning Algorithms** ", International Journal of Emerging Trends in Engineering Research, Volume 8, No. 7, 2020.
21. Mangesh Limbitote , Dnyaneshwari Mahajan , Kedar Damkondwar, Pushkar Patil, 2020, "**A Survey on Prediction Techniques of Heart Disease using Machine Learning**", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 06 (June 2020).
22. Asha Rajkumar, G.Sophia Reena, "**Diagnosis Of Heart Disease Using Datamining Algorithm**", Global Journal of Computer Science and Technology 38 Vol. 10, Issue 10 Ver. 1.0 September 2010.
23. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "**Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction**", IJCSE, Vol. 3, No. 6 June 2011.
24. Aishwarya Mujumdar, V Vaidehi, "**Diabetes Prediction using Machine Learning Algorithms**", Procedia Computer Science, Volume 165, Pages 292-299, ISSN 1877-0509, 2019.
25. N. Nnamoko, A. Hussain and D. England, "**Predicting Diabetes Onset: An Ensemble Supervised Learning Approach**", IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, pp. 1-7, 2018.

26. Huy Pham & Evangelos Triantaphyllou, "**Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization**", Computer and Information Science. SCI, Volume 131, pages 11-26, 2008.
27. T. R. Baitharu and S. K. Pani, "**Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset**", Procedia Computer Sci., vol. 85, no. Cms, pp. 862–870, 2016.
28. Shelly Gupta, Dharminder Kumar and Anand Sharma, "**Data mining classification techniques applied for breast cancer diagnosis and prognosis**", Indian Journal of Computer Science and Engineering, volume 2, 2011.
29. Mehwish Dildar et al., "**Skin Cancer Detection: A Review Using Deep Learning Techniques**", International Journal of Environmental Research and Public Health, Volume 18, No. 10, ISSN 1660-4601, 2021.
30. J. Tang, S. Alelyani, and H. Liu, "**Feature Selection for Classification: A Review**", C. Aggarwal (ed.), Data Classification: Algorithms and Applications. CRC Press, 2014.
31. Y. Dhote, S. Agrawal and A. J. Deen, "**A Survey on Feature Selection Techniques for Internet Traffic Classification**", International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, pp. 1375-1380, 2015.
32. Dr. B. Sarojini, Dr. N.Ramaraj, "**Enhancing Medical Prediction using Feature Selection**", International Journal of Artificial Intelligence & Expert Systems (IJAE), Volume (1) : Issue (3), 2011.
33. Ihsan Abod Khalaf, Abdallah M Abualkishik and Abdulla Amin Aburomman, Mamun Bin IbneReaz, "**Two Features Selection Algorithms based On ensemble of SVM Classifier For Intrusion Detection**", Australian Journal of Basic and Applied Sciences, Volume 7, pp. 480-485, ISSN 1991-8178, 2013.
34. A. Jovic, K. Brkic and N. Bogunovic, "**A review of feature selection methods with applications**", 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200-1205, 2015.
35. Phung, S. L., Bouzerdoum, A., and Nguyen, G. H., "**Learning pattern classification tasks with imbalanced data sets**", P. Yin (Eds.), Pattern recognition, pp. 193-208, 2009.
36. G.M. Weiss and F. Provost, "**The Effect of Class Distribution on Classifier Learning: An Empirical Study**", Technical Report MLTR 43, Dept. of Computer Science, Rutgers Univ., 2001.
37. W.G. Cochran, "**Sampling Techniques. 3rd edn.**" Wiley, New York, 1977.
38. M. Kubat, S. Matwin, "**Addressing the Curse of Imbalanced Training Sets: One-Sided Selection**", Fisher, D.H. (ed.): Proceedings of the Fourteenth International Conference in Machine Learning. Morgan Kaufmann, San Francisco, pp. 179-186, 1997.
39. J. Laurikkala, "**Improving Identification of Difficult Small Classes by Balancing Class Distribution**", Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine, pp. 63-66, 2001.
40. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "**SMOTE: Synthetic Minority Over-sampling Technique**", Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002.
41. S. F. Abdoh, M. Abo Rizka and F. A. Maghraby, "**Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques**" IEEE Access, vol. 6, pp. 59475-59485, 2018.