

A KNN approach for COVID-19 diagnosis with the use of Chi-square for Feature selection

^{1,*}Roseline Oluwaseun Ogundokun, ²Oluwakemi Christiana Abikoye, ³Rasheed Gbenga Jimoh, ⁴Peter Ogirima Sadiku, ⁵Joseph Bamidele Awotunde

¹*Department of Computer Science, Landmark University Omu Aran, Nigeria, ogundokun.roseline@lmu.edu.ng*

²*University of Ilorin, Ilorin, Kwara State, Nigeria, abikoye.o@unilorin.edu.ng*

³*University of Ilorin, Ilorin, Kwara State, Nigeria, jimoh_rasheed@unilorin.edu.ng*

⁴*University of Ilorin, Ilorin, Kwara State, Nigeria, sadiku.po@unilorin.edu.ng*

⁵*University of Ilorin, Ilorin, Kwara State, Nigeria, awotunde.jb@unilorin.edu.ng*

*Corresponding Author Details: Roseline Oluwaseun Ogundokun, Landmark University Omu Aran, ogundokun.roseline@lmu.edu.ng

Abstract

The coronavirus disease 2019 epidemic (COVID-19) that arose has led to a key community medical hazard in Wuhan, China. Indicators that could be utilized as best prognosticators for COVID-19 patients' experimental results are crucial to be recognized. Machine learning (ML) is effective in helping to make decisions and forecasts based on the vast volume of data generated by the healthcare sector. The most imaginable approach to improving the probability of patient recovery is early detection of COVID-19 illness. In this article, a novel approach named K-Nearest-Neighbour technique (CHI-KNN) was used to diagnose the stage of patients' illness, for which a chi-square algorithm was used to effectively select features to minimize the dimensions of the dataset and increase the classifier speed. The greatest value for k was calculated utilizing an investigational method to maximize the precision of the projected algorithm. The system was evaluated using the confusion matrix in terms of accuracy, precision, and sensitivity. The system had a performance accuracy of 99.5%. The study was also compared the proposed CHI+KNN algorithm with other previous researches.

Keywords: Machine learning, COVID-19, Chi-square, K-Nearest Neighbour, Classification

1. Introduction

It is an incredibly difficult process to diagnose disease and multiple studies are typically required for patients to draw a detailed conclusion. This will cause us to utilize diagnostic instruments intended to aid physicians in their decisions (Maleki, Zeinali & Niaki, 2021). The early decision reduces the time of recovery and can save lives. All of these disorders are the malignant development of the lungs, which develops as the cells in the lung tissues multiply uncontrollably (Maleki, Zeinali & Niaki, 2021).

In Wuhan, Hubei Province, China (Phelan, Katz & Gostin, 2020), an occurrence of acute respiratory syndrome coronavirus (CoV) pneumonia arose in December 2019 and received intense interest globally (Ogundokun, & Awotunde, 2020; Adeniyi, Awotunde, Ogundokun, Kolawole, Abiodun, & Adeniyi, 2020; Lukman, Rauf, Abiodun, Oludoun, Ayinde, Ogundokun, 2020). The World Health Organization (WHO) identified the 2019 COVID-19 virus based on its recognition from the pharyngeal swab sample of a patient (Xu, Shi, Wang, Zhang, Huang, Zhang, ... & Wang, 2020; World Health Organization, 2020; Daniyal, Ogundokun, Abid, Khan, Ogundokun, 2020; Ogundokun, Lukman, Kibria, Awotunde, Aladeitan, 2020; World Health Organization, 2020). SARS-nCov 2 is a CoV species that belongs to the largest, enveloped, single-stranded, positive-sense RNA virus family (Su et al., 2016). In monitoring the growing severity of the COVID-19 pandemic, the science community and infectious response organizations face immense challenges. The epidemic has spread exponentially across the world by 26 June 2020, with more than 9,621,470 established incidents and 487,295 demises globally and it had infected 213 countries (COVID-19 CoV-Update <https://virusncov.com/>, accessed 26 June 2020). The least absolute regression of the shrinkage and selection operator (LASSO) was utilized to classify major severity transition factors in patients with COVID-19 (Liu et al., 2020), and unsympathetically sick individuals showed respiratory failure, severe breathing distress condition, cardiac failure, and septic tremor, which increased the demise of persons with COVID-19 (Zhou et al., 2020).

Earlier findings have found that elderly persons with asthma, heart disease, chronic breathing ailments, or malignance have an elevated chance of mortality associated with COVID-19 globally (Guan et al., 2020; Huang et al., 2020b; Ji et al., 2020; Wang et al., 2020; Wu & McGoogan, 2020; Zhou et al., 2020). To forecast COVID-19 ailment in an organism, few models have been used, however. To forecast COVID-19 disease built on monotonous test center evaluations, an accurate and robust model is therefore urgently needed. The early diagnosis of badly infected individuals is of excessive importance and will help to maximize the utilization of scarce capital and to ensure appropriate treatment.

Machine learning uses science algorithms to discern trends in extensive databases and to iteratively strengthen this identifiable evidence with additional knowledge. In innumerable submissions, such as commercial, defense, fund, internet-built existence, and falsification detection, medical diagnosis, and some other applications, these algorithms are commonly used (Adebisi, Ogundokun, & Abokhai, 2020; Adeyinka, Adebisi, Akande, Ogundokun, Kayode, & Oladele, 2019; Oladele, Ogundokun, Kayode, Adegun, Adebisi, 2019; Ogundokun & Awotunde, 2020; Adegun, Ogundokun, Adebisi, & Asani, 2020; Awotunde, Ogundokun, Ayo, & Matiluko, 2020; Adegun, Akande, Ogundokun, & Asani, 2018) attaining to diverse sorts of knowledge collected uninterruptedly and over multiple channels. Since medical knowledge is most commonly unavailable for open investigation, it may be a difficult challenge to use these techniques to determine disease outcomes (Lynch et al., 2017).

Quite a few procedures, for instance, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Classification Tree (C4.5), Gradient Boosting Machine (GBM), etc. are used in machine learning. Although all of these approaches process knowledge otherwise, an insufficient lately projected ML techniques in the field of finding tumor growth are reviewed chronologically in this section.

Chen et al., 2013 proposed a Fuzzy method utilizing KNN (FKNN) for analysis of Parkinson's ailment (PD). Besides, to discover the greatest unequal characteristics on which the ideal FKNN procedure was developed, they used the key component analysis. They compared their scheme with the algorithm of the SVM and found that they worked better with their proposed approach. The best precision of their FKNN for classification achieved 96.07 percent.

Odajima & Pawlovsky, 2014 claimed that with the sum of neighbors and the amount of knowledge exploited for classification, the accuracy of the KNN system varies. In the meanwhile, descriptions of the variance of extreme and least accuracy values with the sizes of the classification set and the number of neighbors were shown.

Iyer et al., 2015; Hayashi and Yukita, 2016; Sa'di et al., 2015; and Huang et al., 2015 utilized the Decision Tree (DT) approach and obtained 77%, 84%, 77%, and 62% precision correspondingly.

Lynch et al., 2017 added to the SEER database several supervised learning classification methods for instance Linear Regression (LR), DT, GRM, SVM, and a personalized ensemble to order survival for lung cancer patients. The findings revealed that the GBM with a Root Mean Square Error (RMSE) value of 15.322 was the utmost reliable of the five individual models used.

Hashi et al., 2017 used DT and KNN procedures from the Pima Indians Dataset to diagnose diabetes disorder, using 768 datasets, all with 8 attributes, and achieved 90 percent and 77 percent accuracy correspondingly. This means that, in terms of classification precision in this situation, the DT is the exceedingly supervised technique.

Khateeb and Usman, 2017 utilized NB, KNN, J48, and Bagging classifiers. ML classification procedures on heart ailment data comprising of 303 cases, all with 14 characteristics were utilized. They split their investigational findings into 6 instances and the kNN classifier using all 14 attributes found the maximum precision of 79.20 percent.

Pradeep and Naveen, 2018 utilized SVM, NBs, and C4.5 methods to assist specialists with improved cancer survivability degree assumptions in the North Central Cancer Treatment Group (NCCTG) utilizing Lung Cancer data collection. The findings indicate that C4.5 increases the estimation of lung tumors with an improvement in the training data collection.

Joshi and Mehta, 2018 used a recognized ML procedure named KNN to validate its application on the Wisconsin investigative Breast Cancer Dataset. 569 cases with 32 attributes and 2 classes were involved in the dataset. They used two methods to minimize critical dimensionality Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and found that KNN performed better with LDA than KNN and KNN with PCA with 97 percent, 95 percent, and 96 percent precision correspondingly.

Proceedings of the 2nd International Conference on ICT for National Development and Its Sustainability, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria- 2021.

Akben, 2018 used KNN, SVM and NBs pre-process results (CKD) to diagnose long-lasting kidney ailment. They initially utilized the raw data techniques and pointed out that the description was not sufficiently precise to attract therapeutic physicians. As such, immediately the data was pre-processed by the K-means clustering technique, they employed the methods. The results showed that accuracy was meaningfully improved, particularly for the KNN classifier, which attained 96 percent.

Lakshmananprabu et al., 2019 established an amalgam procedure including an Optimal Deep Neural Network (ODNN) and Linear Discriminant Analysis (LDA) to characterize Lung lumps as either malevolent or benevolent. The ODNN was initially utilized in their work to draw out significant features from Computed Tomography (CT) lung images, then LDA was utilized to decrease the dimensionality of the features. Lastly, to refine the ODNN, an improved gravitational search procedure was used. Their algorithm's sensitivity, precision, and accuracy were shown to be 96.2%, 94.2%, and 94.56%, respectively.

Therefore, to possibly diagnose the COVID-19 disease, we proposed the use of the Chi-square K-Nearest Neighbor Chi-square-KNN) machine-learning algorithm. Chi-square was used for feature selection of the COVID-19 datasets after which the datasets got was passed into the KNN ML classifiers and diagnoses are made based on the reassurance gotten from the implementation and related disease management measures need to be emphasized to defend these patients against SARS-CoV-2.

The remainder of this article is formulated in the following way. Section 2 offers a summary of what has been done in the literature on COVID-19 and which ML algorithms have been used. Section 3 outlines the proposed methods meticulously, and the output of the projected procedure is evaluated in Section 4. Finally, this work is concluded in Section 5 and proposes possible research.

2. Material and Methodology

This section discussed the materials such as dataset and methods such as the machine learning algorithm used for the implementation of this investigation. The ML adopted in this study is Chi-Square and KNN algorithms and they will be discussed in this section also.

2.1. Dataset

The dataset utilized for this investigation was gathered from the NCBI database. The dataset consists of a sample of 234 affected persons with COVID-19 (n=93), other virus-related (n=100), or non-virus-related (n=41) acute respiratory illnesses (ARIs). The datasets were downloaded from the following links:

<https://www.ncbi.nlm.nih.gov/gds/?term=covid>

<https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE156063>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156063>.

2.2. Proposed system

The proposed system was implemented using MATLAB 2015. The system used COVID-19 datasets. The datasets were passed into the feature selection algorithm in this case Chi-square to determine useful values or variables from the datasets passed and also to rank the features according to their relevance. The K-Nearest Neighbour (KNN) procedure which is an ML classification procedure was utilized for disease detection.

2.3. System block diagram

Figure 1 shows the proposed system block diagram. The COVID-19 datasets are loaded into the system after which the Chi-square feature selection algorithm is used on it and then the remaining datasets are passed into the KNN classifier algorithm. The system then classifies the datasets and the results gotten are evaluated.

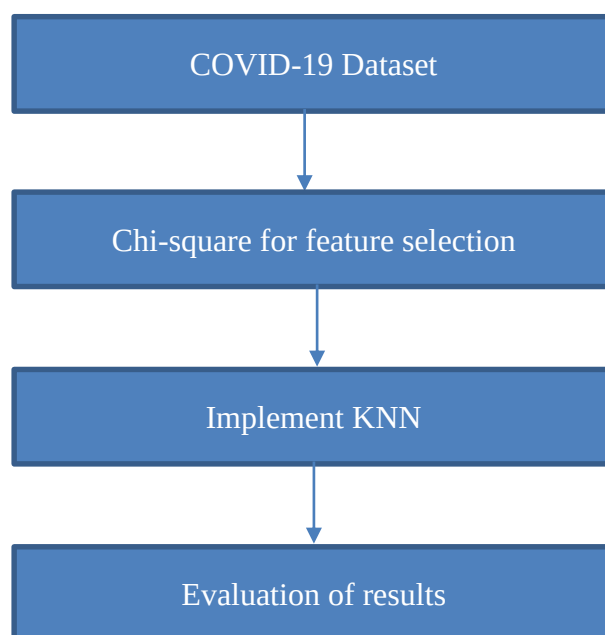


Figure 1. The proposed system block diagram

2.4. Chi-square (χ^2) feature selection algorithm

Chi-squared is an algorithm for the selection of features and has been used for feature selection (Jeung, Baek, Beom, Cho, Her, Yoon, 2019; Wu, Jing, Liu, Ma, Yuan, Wang, et al., 2020) in various research papers. From each feature and target variable, this algorithm calculates χ^2 and selects the required number of features with the best χ^2 scores. If a function is independent of the target variable, the institution is uninformative in classifying observations. For categorical properties in a data set, the Chi-square test is used. Between each feature and the goal, we measure the Chi-square and pick the optimal number of features with the best Chi-square scores. It decides whether the relationship between the two sample categorical variables represents their actual association in the population.

2.5. K-Nearest Neighbour (KNN)

The KNN procedure is a popular ML method built on computer codes to give human-acquired knowledge to run down and create. It is straightforward, spontaneous, quick to execute, and has an extensive variety of applications practically ideal for various kinds of data structures. As a result, the KNN method has been widely used as a non-parametric technique appropriate for non-inactive and non-linear active structures for forecast analysis in various research areas: Davis and Nihan (1991) utilized the KNN technique in interim highway traffic forecasting. Mehrotra and Sharma (2006) proposed a resampled KNN technique optimized for the utilization of manifold prognosticator variables. Bannayan and Hoogenboom (2008) established a method to forecast the understanding of instantaneous regular climate results by incorporating two separate approaches for the KNN method. Mangalova and Agafonov (2014) considered a modeling technique to forecast the power generation of wind farm power generators, and the KNN process was the fundamental modeling solution.

The performance of the KNN procedure is based on the data's demonstrative and comprehensive nature, comparable to other data-driven procedures. As a standard regression method, the KNN procedure has an actual basic operating process: k signifies the sum of the closest neighbors. In the training set, K training models can be initiated, built on the distance extent technique, determine the k value to minimize the error, and the length of the sample, l . In other words, the fundamental postulation of the KNN procedure is that an additional or reduced amount comparable to that projected in the upcoming neighborhood may be initiated in the previous neighborhood. The KNN method does not level the old data, particularly once distinct events transpire, the forecast performance was improved compared to the parametric process.

Proceedings of the 2nd International Conference on ICT for National Development and Its Sustainability, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria- 2021.

3. Results

The two machine learning algorithms were introduced using the kit MATLAB 2015. After the reduced COVID-19 datasets were collected, the COVID-19 datasets were entered into the MATLAB and then moved into the PPSO algorithm. The decreased dataset was then moved to the KNN classifier, and the outcome was collected. The efficiency metrics used for the measurement is the uncertainty matrix, and some interfaces from the design and testing of the framework are below.

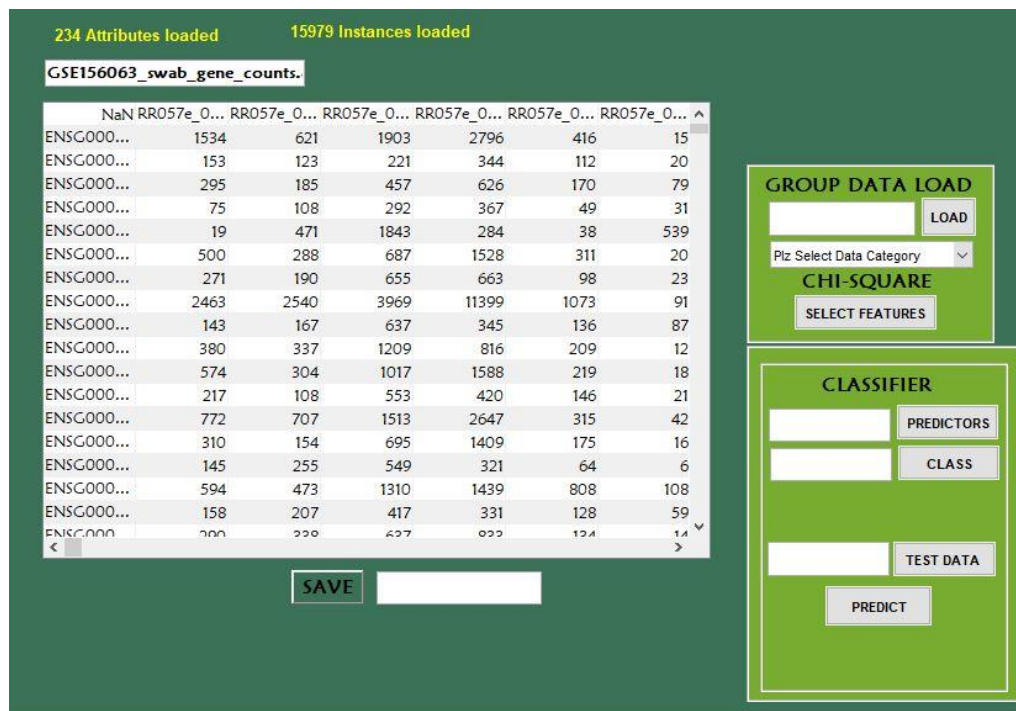


Figure 2. COVID-19 datasets uploaded

The COVID-19 datasets used for the implementation can be seen in Figure 2. The original COVID-19 datasets inserted into the MATLAB are the basic interface in Figure 2 and the Chi-square algorithm was used for the preprocessing point, which is the dataset extraction function.

	CHI-SQUARE	ATTRIBUTES		Clump thickness	Uniformity of cell size	Marginal Adhesion	Single epithelial cell size	Bare nuclei	Bland chomaz
1	542.7759 2		1	4	4	7	10	3	
2	524.8727 3		2	1	1	2	2	3	
3	445.0045 5		3	8	8	3	4	3	
4	495.0569 6		4	1	1	2	1	3	
5	459.4155 7		5	10	10	7	10	9	
6	419.5262 8		6	1	1	2	10	3	
			7	1	2	2	1	3	
			8	1	1	2	1	1	
			9	2	1	2	1	2	
			10	1	1	1	1	3	
			11	1	1	2	1	2	
			12	3	3	2	3	4	
			13	7	5	7	9	5	
			14	4	6	6	1	4	
			15	1	1	2	1	2	
			16	1	1	2	1	3	
			17	7	7	4	10	4	
			18	1	1	2	1	3	
			19	3	2	5	10	5	
			20	5	5	6	7	7	
			21	1	1	2	1	2	
			22	4	5	2	0	7	
			23	1	1	2	1	3	
			24	2	3	2	7	3	
			25	2	1	1	1	2	
			26	1	1	2	1	2	

Figure 3. Reduced Dataset using Chi-square algorithm

Figure 3 shows the reduced COVID-19 datasets using the Chi-square algorithm.

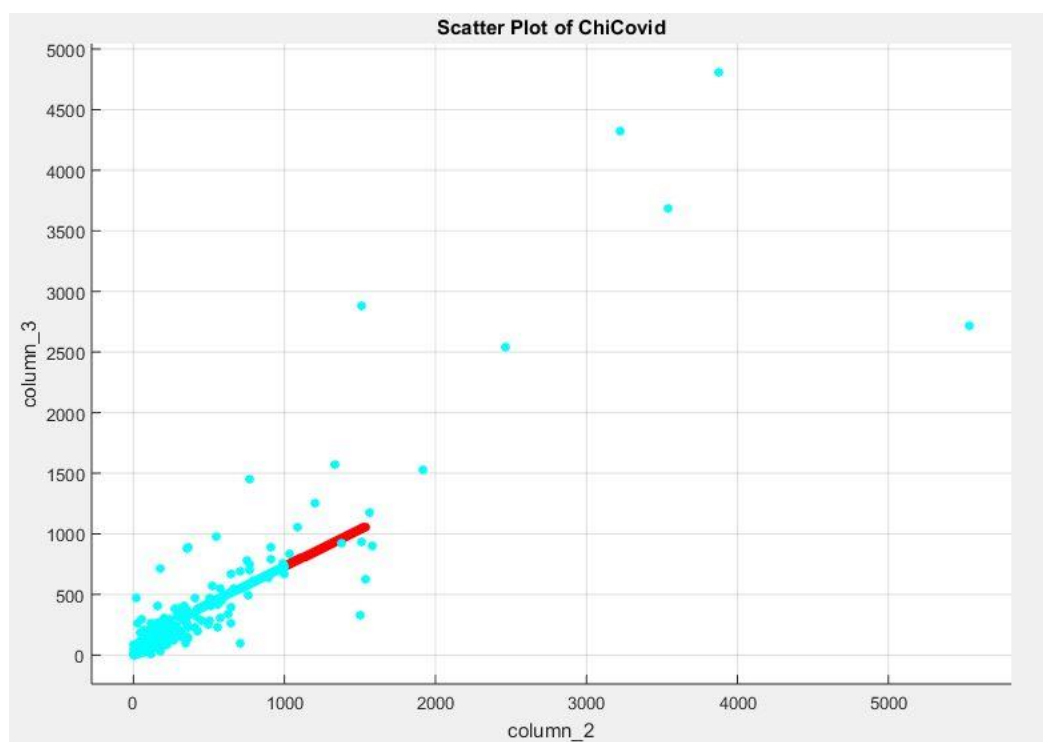


Figure 4. Scattered plot for the KNN

Figure 4 displays a dispersed K Nearest Neighbour map displaying the COVID-19 dataset classification for disease identification and prediction.

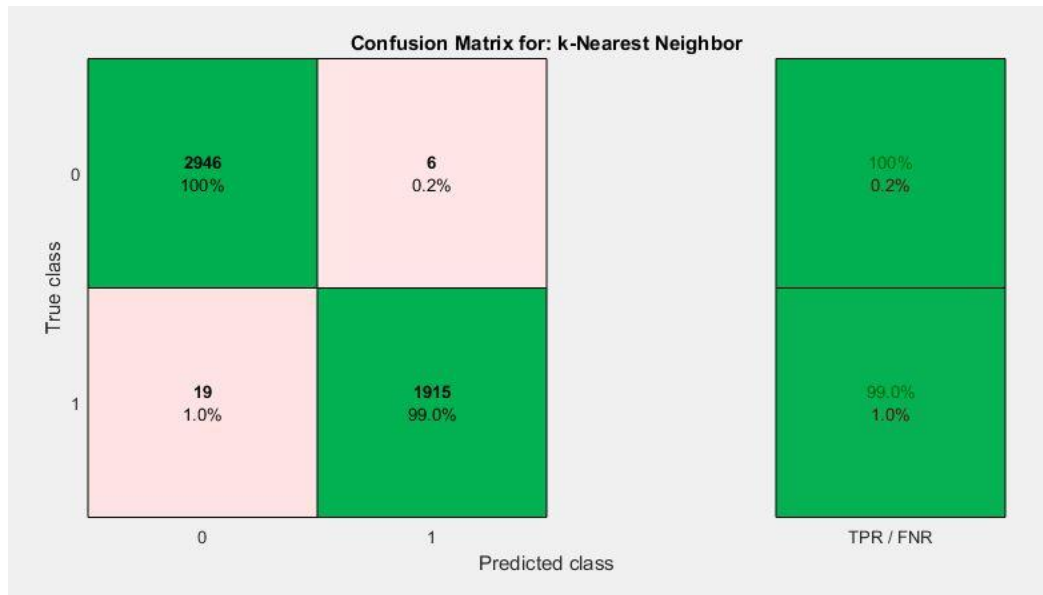


Figure 5. Confusion matrix for Chi-square with KNN (CHI+KNN)

For the proposed method, Figure 5 displays the Chi-square and K Nearest Neighbor (KNN) confusion matrix.

3.1 K-Nearest Neighbour (KNN)

Using the confusion matrix, the established method was tested. Sensitivity, precision, specificity, accuracy, F1 ranking, negative predictive value, false-positive rate, false discovery rate, false-negative rate, and Matthews correlation coefficient were the measurements used in this analysis.

The performance evaluation metrics of Chi-Square and KNN is calculated as follows:

$$\text{Correctly Classified Instances (accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

For this system, the TP, TN, FP, and FN values are as follows

TP = 2946

TN = 1915

FP = 6

FN = 19

Table 2: Confusion matrix for proposed system evaluation

Measures	Value
Sensitivity	0.9936
Specificity	0.9969
Precision	0.9980
Accuracy	0.9949
F1 Score	0.9958

Negative Predictive Value	0.9902
False Positive Rate	0.0031
False Discovery Rate	0.0020
False Negative Rate	0.0064
Mathews correlation coefficient	0.9893

Table 2 displays the confusion matrix for the proposed assessment of the device, and it was shown that the system had 99.8% precision, 99.4% sensitivity, 99.7% specificity, 99.5% accuracy, and 99.6% F1 score. This has demonstrated that the method is an accurate and efficient one.

3.2 System Comparison

The system was compared with previous researches that had used machine learning algorithms, as shown in Table 3 below

Table 3. System comparison with previous researches

Authors	ML Methods used	Disease predicted	Accuracy	Sensitivity	Specificity	Precision
Lakshmanaprabu et al., 2019	Optimal deep neural network (ODNN) and linear classification analysis (LDA)	Lung Images	94.6%	96.2%	-----	94.2%
Babaoglu, Fundik, Ülker, & Aygül, 2012	AIRS, PSO, and KNN	Coronary artery disease	76%	86%	54%	82%
Proposed System	Chi-square and KNN	COVID-19	99.5%	99.4%	99.7%	99.8%

4. Discussion

Compared with three previous pieces of analysis, the scheme was found to work better in terms of accuracy, sensitivity, specificity, and precision in the proposed review. Lakshmanaprabu et al., 2019 used two ML algorithm which is an optimal deep neural network (ODNN) and linear classification analysis (LDA) for lung cancer prediction, and got a 94.6% of accuracy, 96.2% sensitivity, and 94.2% precision. Babaoglu, Fundik, Ulker & Aygul, 2012 used three ML algorithms to predict coronary disease and had 76% accuracy, 86% sensitivity, 54% specificity, and 82% precision. But our proposed study used a Chi-square algorithm on the COVID-19 datasets for feature selection with KNN for prediction and monitoring and had an accuracy of 99.5 percent, 99.4 percent sensitivity, 99.8 percent precision, and 99.4 percent specificity, which indicated that our proposed study worked higher, efficiently and effectively.

5. Conclusions

Machine learning is a popular approach for predictive modelling, yet it does not play a significant role in combating COVID-19. Machine learning can help define treatment effects, pharmaceutical views, diagnostic prediction, and improve several facets such as the economy the recent worldwide pandemic can bring. Improving proposed approaches with its risk factors should concern and present efficient algorithms for rapid prediction, and improvement should be considered. In this study, Chi-square and KNN (CHI+KNN) was used for predicting the novel Covid-19, and an accuracy of about 99.5% was attained. Health care has been assisted by frameworks to evaluate the strength of suggestions over the years. Several algorithms for assessing and understanding machine

learning have been proposed, and significant public categories of learners can help predict and analyse COVID-19. In the future, several ML algorithms could be utilized to enhance the performance and compared with the state-of-art, using other approaches such as probabilistic PCA, Probabilistic Partial Least Square with classifiers such as DT, SVM, ANN, RF, and so on.

Acknowledgement

Acknowledgement of financial support can be stated here. Any other technical assistance rendered by any agency/body/individual may be acknowledged.

References and citation

All references and citations should follow APA 6.7 format. In the text, a work is cited by means of author's surname followed by year of publication in parentheses. For example, software engineering projects are hardly completed on schedule and within budget unless good software project management techniques are enforced (Oladele & Mojeed, 2014). Akula and Cusick (2008) carried out a statistical survey on four real-life software projects to analyse the impact of overtime on software quality. With respect to error propagation during software development projects, Jones(2000) found out that developers introduced an average of 4 errors/FP through analysis, design and coding activities of a component. Recently, Sarro et al. (2017) extended the work of Ferrucci et al. (2013) by proposing an adaptive method for meta-heuristic operator selection to improve algorithmic performance for software overtime planning problem. All cited works are listed in alphabetical order in the reference list.

References

- Akula, B. & Cusick, J. (2008). Impact of overtime and stress on software quality. *Proceedings of the 4th International Symposium on Management, Engineering, and Informatics (MEI)*. Orlando, Florida, United States of America, 9-18.
- Maleki, N., Zeinali, Y., & Niaki, S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, 164, 113981.
- Ogundokun, R. O., & Awotunde, J. B. (2020). Machine Learning Prediction for COVID 19 Pandemic in India. *medRxiv*.
- Adeniyi, E. A., Awotunde, J. B., Ogundokun, R. O., Kolawole, P. O., Abiodun, M. K., & Adeniyi, A. A. (2020). Mobile health application and COVID-19: Opportunities and challenges. *Journal of Critical Reviews*, 7(15), 3481-3488.
- Lukman, A.F., Rauf, R.I., Abiodun, O., Oludoun, O., Ayinde, K., Ogundokun, R.O. (2020). COVID-19 prevalence estimation: Four most affected African countries. *Infectious Disease Modelling*, 5, pp. 827-838
- Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., ... & Wang, F. S. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet respiratory medicine*, 8(4), 420-422.
- World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Date Accessed 17/06/2020.
- Ogundokun, R.O., Lukman, A.F., Kibria, G.B.M., Awotunde, J.B., Aladeitan, B.B. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, 5, pp. 543-548
- Daniyal, M., Ogundokun, R.O., Abid, K., Khan, M.D., Ogundokun, O.E. (2020). Predictive modeling of COVID-19 death cases in Pakistan. *Infectious Disease Modelling*, 5, pp. 897-904
- World Health Organization. <https://covid19.who.int/>. Date Accessed 17/06/2020.
- Su, P. H., Gasic, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., ... & Young, S. (2016). Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

- Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*, 395(10229):1054–1062. DOI 10.1016/S0140-6736(20)30566-3.
- Guan W-J, Ni Z-Y, Hu Y, Liang W-H, Ou C-Q, He J-X, Liu L, Shan H, Lei C-L, Hui DSC, Du B, Li L-J, Zeng G, Yuen K-Y, Chen R-C, Tang C-L, Wang T, Chen P-Y, Xiang J, Li S-Y, Wang J-L, Liang Z-J, Peng Y-X, Wei L, Liu Y, Hu Y-H, Peng P, Wang J-M, Liu J-Y, Chen Z, Li G, Zheng Z-J, Qiu S-Q, Luo J, Ye C-J, Zhu S-Y, Zhong N-S. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *New England Journal of Medicine*, 382(18):1708–1720. DOI 10.1056/NEJMoa2002032.
- Huang R, Zhu L, Xue L, Liu L, Yan X, Wang J, Zhang B, Xu T, Ji F, Zhao Y, Cheng J, Wang Y, Shao H, Hong S, Cao Q, Li C, Zhao X-A, Zou L, Sang D, Zhao H, Guan X, Chen X, Shan C, Xia J, Chen Y, Yan X, Wei J, Zhu C, Wu C, Santiago Hda C. (2020b). Clinical findings of patients with coronavirus disease 2019 in Jiangsu province, China: a retrospective, multi-center study. *PLOS Neglected Tropical Diseases*, 14(5): e0008280. DOI 10.1371/journal.pntd.0008280
- Ji, H-L., Zhao, R., Matalon, S., & Matthay, M. A. (2020). Elevated plasmin (ogen) as a common risk factor for COVID-19 susceptibility. *Physiological Reviews*, 100(3):1065–1075. DOI 10.1152/physrev.00013.2020
- Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., & Peng, Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA*, 323(11):1061–1069. DOI 10.1001/jama.2020.1585.
- Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72,314 cases from the Chinese center for disease control and prevention. *Jama* 323(13):1239–1242. DOI 10.1001/jama.2020.2648.
- Adebisi, M.O., Ogundokun, R.O., Abokhai, A.A. (2020). Machine Learning-Based Predictive Farmland Optimization and Crop Monitoring System. *Scientifica* 2020
- Adeyinka, A.A., Adebisi, M.O., Akande, N.O., Ogundokun, R.O., Kayode, A.A., Oladele, T.O. (2019). A Deep Convolutional Encoder-Decoder Architecture for Retinal Blood Vessels Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11623 LNCS, pp. 180-189
- Oladele, T.O., Ogundokun, R.O., Kayode, A.A., Adegun, A.A., Adebisi, M.O. (2019). Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11623 LNCS, pp. 716-730
- Adegun, A. A., Akande, N. O., Ogundokun, R. O., & Asani, E. O. (2018). Image segmentation and classification of large-scale satellite imagery for land use: a review of the state of the arts. *Int. J. Civ. Eng. Technol*, 9(11), 1534-1541.
- Adegun, A. A., Ogundokun, R. O., Adebisi, M. O., & Asani, E. O. (2020). CAD-Based Machine Learning Project for Reducing Human-Factor-Related Errors in Medical Image Analysis. *In Handbook of Research on the Role of Human Factors in IT Project Management* (pp. 164-172). IGI Global.
- Awotunde, J. B., Ogundokun, R. O., Ayo, F. E., & Matiluko, O. E. (2020). Speech Segregation in Background Noise Based on Deep Learning. *IEEE Access*, 8, 169568-169575.
- Chen, H.-L., Huang, C.-C., Yu, X.-G., Xu, X., Sun, X., Wang, G., & Wang, S.-J. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. *Expert systems with applications*, 40(1), 263–271.
- Odajima, K., Pawlovsky, A. P. (2014). A detailed description of the use of the KNN method for breast cancer diagnosis. *In Proceedings of the 7th international conference on biomedical engineering and informatics (BMEI)*, Dalian, China.

- Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, 5(1), 1–14.
- Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., ... Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, 108, 1–8.
- Hashi, E. K., Zaman, S. U., Hasan, R. (2017). An expert clinical decision support system to predict disease using classification techniques. *In The proceedings of the 2017 international conference on electrical, computer and communication engineering (ECCE), Cox's Bazar, Bangladesh*
- Khateeb, N., Usman, M. (2017). Efficient heart disease prediction system using K-nearest neighbor classification technique. *In Proceedings of BDIOT2017 proceedings of the international conference on big data and internet of thing, London, UK* (pp. 21–26).
- Pradeep, K. R., & Naveen, N. C. (2018). Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and naive Bayes algorithms for healthcare analytics. *Procedia Computer Science*, 132, 412–420.
- Joshi, A., & Mehta, A. (2018). Analysis of K-nearest neighbor technique for breast cancer disease classification. *International Journal of Recent Scientific Research*, 9(4), 26126–26130.
- Akben, S. B. (2018). Early-stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. *IRBM*, 39(5), 353–358.
- Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, 92, 374–382.
- <https://www.ncbi.nlm.nih.gov/gds/?term=covid>
- <https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE156063>
- <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156063>
- Jeung M, Baek S, Beom J, Cho KH, Her Y, Yoon K. (2019). Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *J Hydrol*; 575:1099–110. doi: 10.1016/j.jhydrol.2019.05.079.
- Chen G, Wang Y, Li S, Cao W, Ren H, Knibbs LD, et al. (2018). Spatiotemporal patterns of PM 10 concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. *Environ Pollut*; 242:605–13. doi: 10.1016/j.envpol.2018.07.012.
- Wu Z, & McGoogan JM. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72,314 cases from the Chinese center for disease control and prevention. *Jama* 323(13):1239–1242. DOI 10.1001/jama.2020.2648
- Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, Zhao Y, Li Y, Wang X, Peng Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA*, 323(11):1061–1069. DOI 10.1001/jama.2020.1585.
- Davis, G. A., & Nihan, N. L. (1991). Nonparametric regression and short-term freeway traffic forecasting. *Journal of Transportation Engineering*, 117(2), 178–188.
- Mehrotra, R., & Sharma, A. (2006). Conditional resampling of hydrologic time series using multiple predictor variables: A K-nearest neighbour approach. *Advances in water resources*, 29(7), 987–999.
- Bannayan, M., & Hoogenboom, G. (2008). Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach. *Environmental Modelling & Software*, 23(6), 703–713.

Mangalova, E., & Agafonov, E. (2014). Wind power forecasting using the k-nearest neighbors' algorithm. *International Journal of Forecasting*, 30(2), 402-406.

Babaoglu, I., Findik, O., Ülker, E., & Aygöl, N. (2012). A novel hybrid classification method with particle swarm optimization and k-nearest neighbor algorithm for diagnosis of coronary artery disease using exercise stress test data. *International Journal of Innovative Computing, Information and Control*, 8(5), 3467-3475.

Authors Biography



Roseline Oluwaseun Ogundokun is currently a Lecturer at the Department of Computer Science, College of Pure and Applied Sciences, Landmark University, Omu Aran, Kwara State, Nigeria, she holds Bachelor of Science in Management Information System from Covenant University, Ota; Master of Science in Computer Science from the University of Ilorin, Ilorin; Post Graduate Diploma in Education (PGDE) from the National Teachers' Institute (NTI), Kaduna and; currently a PhD student in the Department of Computer Science, University of Ilorin, Ilorin. Her research interests include Machine learning, Deep Learning, Steganography and Cryptography, Information Security, Artificial Intelligence, Data Mining, Information Science, and Human Computer Interaction.



Dr. Oluwakemi Christiana Abikoye is currently an Associate Professor at the Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin. She received her B.Sc., M.Sc. and Ph.D. Degrees in Department of Computer Science. Her research interest includes Cryptography, Computer and Communication Network (Cyber) Security, Biometrics, Human Computer Interaction and Text and Data Mining.



Prof. Rasheed Gbenga Jimoh is currently a Professor at the Department of Computer Science, University of Ilorin, Ilorin, Nigeria, and currently the Dean, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria. His research interests include Human Computer Interaction, Information Security, Knowledge Management, Bioinformatics, and Artificial Intelligence. He is a member of the Academia in Information Technology Profession (MAITP), Computer Professional Registration Council of Nigeria (MCPN), the Institute of Electrical and Electronics Engineers (MIEEE), and the Nigeria Computer Society (MNCS).



Dr. Joseph Bamidele Awotunde is currently a Lecturer with the Department of Computer Science, University of Ilorin, Nigeria. He received the B.Sc. degree in Mathematics/Computer Science from Federal University of Technology, Minna, Nigeria, in 2007. M.Sc. and Ph.D. degrees in Computer Science from the University of Ilorin, Ilorin, Nigeria, in 2014 and 2019 respectively. His research interests include information security, bioinformatics, artificial intelligence, software engineering, and biometrics. He is a member of the International Association of Engineers and Computer Scientist (MIAENG), the Computer Professional Registration Council of Nigeria (MCPN), and the Nigeria Computer Society (MNCS).