International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 1, Jan-Feb 2023

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Hybrid PSO Model for predicting Mortality Risk Among Covid-19 Patients Using SVM Classifier

Muhammad Idris[1,5]; Zahraddeen Sufyanu[2]; Shamsuddeen M. Abubakar[3]; Abdullahi Sani Dauda[4]

[1]Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria
Email: muhammadidrisfud@gmail.com
[2]Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria.
Email: sufyanzzzzz@gmail.com
[3]Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria.
Email: Salsabil012@gmail.com
[4]Department of Computer Science, Federal University of Kashere, Gombe State, Nigeria.
Email: daudasaniaa@fukashere.edu.ng
[5]Department of Computer Science, Binyaminu Usman Polytechnic Hadejia, Jigawa State, Nigeria.

----------**********----------

Abstract:

The Chinese government alerted the rest of the world in December 2019 that a new breed of coronavirus, severe acute respiratory syndrome-related coronavirus (COVID-19), was rapidly spreading across China and had entered other countries. The COVID-19 pandemic has raised many challenges for healthcare professionals, including the need for accurate predictions of the mortality risk among individuals infected with the virus. In this paper, a hybrid particle swarm optimization and support vector machine (PSO-SVM) model is proposed to predict patients with high mortality risk at the early stage of admission to assist medical professionals in making informed decisions regarding the allocation of medical resources during critical situations, thereby safeguarding human lives. The model was trained and validated using a dataset of COVID-19 patient comorbidities and clinical, demographic, and laboratory admission values that were collected using a software tool for healthcare monitoring developed by Streamline Health of Atlanta, Georgia, and a comprehensive review of medical records. The proposed model, PSO-SVM, outperformed SVM with an accuracy of 82.9268% and a ROC AUC of 0.8666. The study provides a promising approach for predicting the mortality risk among COVID-19 patients and may help healthcare professionals make more informed decisions about patient care.

Keywords **—covid-19, particle swamp optimization, support vector machines, machine learning classifier.**

----------**********----------

I. INTRODUCTION

The Chinese government alerted the rest of the world in December 2019 that a new breed of coronavirus, severe acute respiratory syndrome-related coronavirus (COVID-19), was rapidly spreading across China and had entered other countries. The incident was traced back to a seafood market in Wuhan, according to the US Centers for Disease Control and Prevention (CDC). On January 13, 2020, the World Health Organization (WHO) reported a case in Thailand, the first case outside of China. Japan reported its initial case, On January 16 and January 20 South Korea reported their first confirmed case. Currently, the virus has spread to almost every country on the globe [1]. Predicting the spread of COVID-19 disease and examining its epidemiological characteristics are critical problems that should be thoroughly investigated to reduce outbreak prevalence, manage production operations, and allocate medical resources. To anticipate the spread of the pandemic in different nations, statistical and mathematical modelling tools have been applied furthermore, the COVID-19

outbreak has prompted researchers to propose advance statistical and mathematical modelling techniques to predict the spread of the virus in different countries. [2]. Due to nonlinearity of COVID-19 data necessitates the use of multiple machine learning techniques to build a robust forecasting model.However, the principal contribution of this work is an improvement in the performance of Support Vector Machine (SVM) for predicting high mortality risk in COVID-19 patients. Specifically, the study utilizes Particle Swarm Optimization (PSO) for feature selection, which reduces the computational cost and improved the accuracy of the SVM model. Furthermore, this research work aims to assist government efforts to combat the virus by providing healthcare professionals with necessary information about patients who have a high mortality risk. This information can guide decisions on allocating medical resources during critical situations, ultimately helping to protect human lives. In addition, this research evaluates the performance of the proposed model against that of the standard SVM. The remaining sections of the paper are structured as follows: Section II present review of related work, section III presents a detailed description of the dataset and methodology employed in the study, while Section IV presents the results. The results are subsequently discussed in Section V, and Section VI provides concluding remarks and outlines potential avenues for future research.

II. LITERATURE REVIEW

Hybrid models of various intelligence techniques have been commonly utilized in forecasting. [3]. Researchers have identified gaps in the research area and advocate for the development of new forecasting models for predicting future occurrences of the pandemic and Nature-inspired algorithms have shown to be the best approaches for dealing with such problems because of their reliability and effectiveness[4]. In a study conducted by [5] presented a hybrid deep learning framework that outperformed other state-of-the-art classification algorithms for the classification of pneumonia infection in COVID-19 patients using normal chest CT scans. The finding of the research shows that the proposed framework outperformed the other algorithms, achieving an accuracy of 88.7%. however, some important features were removed during feature selection which might affect the model accuracy and can limit the generalizability of the result, [6] also, highlighted that the recent focus of COVID-19 research has been on the application of machine learning algorithms in clinical and laboratory data for the prediction of mortality risk and the diagnosis of COVID-19 patients. Moreover, the study also pointed that supervised learning algorithms are the most used machine learning method in predicting mortality risk and diagnosing COVID-19 patients. However, most proposed models are yet to be implemented in real-world settings. Also, researchers advocate for the use of artificial intelligence and machine learning techniques to improve diagnosis, prognosis, monitoring, and administration of COVID-19 treatments. To mitigate and reduce COVID-19 spread, [7] applied Support Vector Regression and person correlation techniques to predict and analyze COVID-19 spread across different countries and regions, while also correlating COVID-19 transmission with whether conditions and how long it will take for the pandemic to end. The performance of the proposed model was evaluated against other regression models and has shown promising performance compared to other popular regression techniques. [8]created a model using XGBoost Classifier for prediction of patient in urgent need of mechanical ventilation within one day of their initial encounter at the hospital, Respiratory Decomposition (READY) using clinical data, a machine learning technique and evaluated its performance against the old early warning system, the modified early warning score (MEWS). In order to gain insights into the patterns of covid-19 among infected patients, [9] proposed a machine learning technique to predict the patient outcome during hospitalization period, specifically, the probability of survival. To help in early monitoring and effective treatment of infected COVID-19 patients, [10] proposed a hybrid decision level fusion with gradient boosting (GB), random forest (RF), and extreme gradient boosting (XGB) for predicting the outcome of the infected patient (recovery or death) using clinical, geographical, and demographic data. The results shows that these algorithms performed exceptionally well in classifying COVID-19 patients, with an accuracy of over 91%. However, further investigation is needed in identifying areas that may have clusters of the virus, evaluating the effects of the virus on pregnant patients and those with chronic or long-standing health conditions. Moreover, there is a critical need to validate the classifiers used in the study on a larger and more reliable dataset for improved results and accuracy.[11] proposes a classifier for predicting the recovery or death status of COVID-19 patients in South Korea using an Artificial Neural Network (ANN) with a single hidden layer and gradient descent as the optimization algorithm. The model performs exceptionally well in classifying death versus recovery cases. Furthermore, it was discovered that the top three variables to predict death status are infection reason, confirmation date, and region. The results indicate that combining the most effective categorical variable with a numerical variable could enhance the performance of the model, nevertheless the study is limited to dataset from south Korea and did not consider other factors that could affect the recovery and death status of COVID-19 patients which make it unable to generalized the result to other population also perform low in classifying recovered than death cases [12] investigate the correlation between the severe symptoms of COVID-19 and comorbidities as well as mortality through a meta-analysis of the existing state of art literature and employed machine learning approaches such as, XGBoost (XGB), Random Forest, Decision Tree, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and Light Gradient Boosting Machine (LGBM) on a compiled reported COVID-19 dataset. The results of the meta-analysis indicated that there is a strong relationship between severe COVID-19 symptoms and several comorbidities however, due to limited data, the research couldn't clearly answer questions

on the relationship between gender and age-related comorbidity. [13] created a model for improving the survival rate of COVID-19 patients by training multipurpose algorithms using a combination of laboratory, clinical, and demographic data. The proposed model demonstrates the potential for using machine learning algorithms to enhance the management and treatment of COVID-19 patients, with the goal of improving their survival rates. It was clearly shown from the state of art that accurate prediction of mortality risk among covid-19 patients can be achieved with machine learning technique applied on important features of covid-19 confirm positive patient. In research conducted by [14] utilized ensemble learning to predict the severity of COVID-19 patients in the early stages by using information about their admission laboratory values, demographics, comorbidities, admission medications, admission supplementary oxygen orders, discharge, and mortality. The study tries out 17 different machine learning models and a voting classifier, which is an ensemble of the top models, on a dataset of 4711 patients with 85 features. The results show a best AUC of 0.89. However, this research employs single algorithm for feature selection which reduces the features from 85 to 43.

III. MATERIALS AND METHODOLOGIES

A. *Data Description*

This work was conducted on the dataset used in [14] which contains diverse covid-19 patient information such as admission laboratory values, demographics, comorbidities, admission medications, admission supplementary oxygen orders, discharge, and mortality. The data was collected using a healthcare monitoring software tool named Clinical Looking Glass (CLG) by Streamline Health based in Atlanta, Georgia, and a significant medical record review. The dataset consist of 4,711 records of patients diagnose with Covid-19 infection which includes 85 variables, including length of hospital stay (LOS), myocardial infarction (MI), cardiovascular disease (CVD), diabetes mellitus simple (DM simple), diabetes mellitus complicated (DM complicated), congestive heart failure (CHF), dementia (Dement), Chronic obstructive pulmonary disease (COPD), oxygen saturation (OsSats), peripheral vascular disease (PVD), mean arterial pressure in mmHg (MAP), D-dimer in mg/ml (Ddimer), platelets in k per mm3 (Plts), international normalized ratio (INR), blood urea nitrogen in mg/dL (BUN), alanine aminotransferase in U/liter (AST), white blood cells in per mm3 (WBC), and interleukin-6 in pg/ml (IL-6).

B. *Data Pre-processing*

This section presents a comprehensive description of the how the data was preprocess. The initial task involved in developing the prediction model was data pre-processing and exploratory data analysis to gain insights into the dataset. Although the dataset did not contain any missing values, it contained outliers, and the age variable was of type string. Consequently, one-hot encoding was utilized to convert it into an integer. Furthermore, the derivation cohort column was dropped since it was employed to partition the dataset into two for cross-validation purposes. Subsequently, the distribution of each column in the dataset was examined, and relationships between each column and our target, which is death, were identified. Outliers were then detected and removed, and the clean dataset was saved in a new location.

C. *Support Vector Machine and Particle Swarm Optimization*

Support Vector Machine

SVM is a powerful machine learning technique used for classification, regression, and outlier detection. Its basic principle involves transforming a training dataset into a higher-dimensional space, where it optimizes a hyperplane to determine the hyperplane that best separates the data into various classes with the lowest classification error rate [15]. Below is a representation of the hyperplane:

$$W.X + b = 0 \quad (1)$$

In which, W denote the weight vector and b is a scalar representing bias.

Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based optimization algorithm inspired by the collective behavior of animals in nature. Introduced in 1995 by Kennedy and Eberhart, it involves a set of particles moving through a multidimensional search space to find the optimal solution to a given problem [16].

Each particle in the search space maintains a position and velocity, respectively represented by the vectors x and v. The position vector x represents the current solution to the problem, whereas the velocity vector v represents the direction and velocity of the particle's motion. Initially, PSO was developed for continuous optimization problems, but later, Binary PSO was developed for binary optimization problems, such as feature selection (BPSO)

D. Modelling

This section provide description of how the model was created; the dataset was partitioned into two segments. The first segment comprised selecting the features and second our target, which is death. Also, the dataset was divided into two portions; wherein 80% was designated for training and 20% for validation. A robust scaler was used to ensure that the dataset conformed to a single scale. This scaling algorithm was employed since the values in the dataset were of different scales. A support vector machine (SVM) was then created as the first model using 10-fold cross-validation, followed by using random search. Lastly, particle swarm optimization (PSO) was utilized to select the optimal features for running the SVM model. PSO selected the best 43 features out of the initial 85 features, and the SVM model was run using the best 43 features selected by PSO and 10-fold cross-validation. Table I present the hyperparameters of the SVM.

TABLE I
HYPERPARAMETERS OF SVM MODEL.

C	Kernel
0.1, 1, 10	Linear, rbf, poly

E. ***Evaluation Metrics***

Precision, recall, and F1-score are some of the metrics that can be used to evaluate the performance of classification models in machine learning. These metrics are derived from the confusion matrix, a table displaying the number of true positives, false positives, true negatives, and false negatives.

Precision is the ratio of accurate positive predictions to total positive predictions. Calculated as the ratio of true positives (TP) which implies that the patient is at high risk of mortality to the sum of true positives and false positives (FP) which implies the patient is not at high risk for mortality, it generally measures the accuracy of positive predictions. Greater precision implies a reduces number of false positives. Those term TP and FP are defined in same manner for accuracy, recall and f1-score.

It can be represented as

$$Precision = T_P/(T_p + F_P) \quad (2)$$

Recall, also referred to as sensitivity or true positive rate, is the proportion of accurate positive predictions relative to actual positive instances. Calculated as the ratio of true positives to the sum of true positives and false negatives, this metric evaluates the model's ability to identify all positive instances. The greater the recall, the fewer false negatives.

It can be represented by the formular as

$$Recall = T_P/(T_p + F_P) \quad (3)$$

The F1-score is the harmonic mean of precision and recall and a measure of a classifier's overall performance.

It is represented by the formula.

$$F_1 = (2 * Prec * Recall)/(Prec + Recall) \quad (4)$$

F1-score is advantageous when the dataset is unbalanced because it provides a more balanced view of the model's performance than accuracy alone.

Accuracy is calculated as the ratio of true positives and true negatives to the total number of instances and represents the proportion of correct predictions among all predictions. It is a common evaluation metric, but it can be misleading in datasets where the majority class dominates the prediction.

It can be represented as

$$Accuracy = (T_p + T_n)/(T_p + F_p + T_n + F_n) \quad (5)$$

The receiver operating characteristic area under the curve (ROC AUC) quantifies the classifier's ability to distinguish between positive and negative classes. It is the area under the curve generated by plotting the true positive rate versus the false positive rate for various threshold values. The AUC is a numeric value that ranges between 0.5 and 1; a higher AUC value close to 1 indicates a more effective classifier; however, in this study model comparison is based on accuracy and AUC score.

IV. RESULT

In this part, we present and compare the performance of the traditional Support Vector Machine (SVM) and the Particle Swarm Optimization-based SVM (PSO-SVM) using several evaluation metrics. The assessment of the models was based on their accuracy and area under the curve (AUC) score. Furthermore, to compare the performance of our Support Vector Machine (SVM) model with that of (Walia &Jeevaraj, (2021), 10-fold cross-validation was employed. The dataset was divided into 3768 samples for training and 943 for testing, and Table I presents the result of the prediction model. Our primary objective was to predict high-risk patients, and the SVM model attained a ROC AUC of 0.7529 with an accuracy of 80.69 percent. With an accuracy of 82.9268 percent and a ROC AUC of 0.8166, the PSO-SVM model outperforms the SVM model. Results indicate that combining PSO with SVM can enhance model performance. Additional evaluation metrics, such as precision, recall, and f1-score, were also used and are presented in Table II.

TABLE III
RESULT COMPARISON BETWEEN SVM AND PSO-SVM USING 80/20 TRAIN TEST SPLIT.

Algorithm	ROC AUC	Accuracy	Precision	Recall	F1-score
SVM	0.7529	80.6999%	0.82	0.80	0.80
PSO-SVM	0.8166	82.9268%	0.82	0.80	0.80

V. DISCUSSION

Based on the analysis conducted, the findings indicate that patients diagnosed with COVID-19 and presenting with co-morbidities such as seizure, stroke, renal disease, COPD, MI, and CHF have a significantly higher risk of mortality. Additionally, it was observed that the removal of outliers in Length of Stay (LoS) led to an improvement in the performance of the model.

Moreover, the Support Vector Machine (SVM) model developed in this study using 10-fold cross-validation demonstrated good performance with an accuracy of 85% and receiver operating characteristic (ROC) area under the curve (AUC) of 0.8622. This performance is notably better than that of [14], who reported an accuracy of 81% and an ROC AUC of 0.6540. The result of the comparison is presented in Table III.

TABLE IIIII
COMPARISON OF OUR WORK WITH PREVIOUS WORK

Author	Year	Method	ROC_AUC	Accuracy
[14].	2021	SVM	0.6540	81%
Our work	2023	SVM	0.8622	85.0321%

VI. CONCLUSIONS

This research work presents a prediction model that utilizes a hybrid machine learning technique, Particle Swarm Optimization-Support Vector Machine (PSO-SVM), to effectively identify patients with a high risk of mortality at an early stage of admission. This model can assist medical professionals in making informed decisions regarding the allocation of medical resources during critical situations, thereby safeguarding human lives. Furthermore, the performance of the proposed model is evaluated against that of the standard SVM, and the results demonstrate that the newly created SVM model outperforms previous studies with an accuracy of 85% and ROC AUC of 0.8622, as determined through 10-fold cross-validation. Additionally, the performance of the proposed PSO-SVM is compared with that of the standard SVM, using an 80:20 train split, and the PSO-SVM model is shown to outperform the standard SVM with an accuracy of 82.93% and ROC AUC of 0.8166. Further research efforts should explore the application of multi-level machine learning techniques and deep learning to enhance the performance of the model using a more reliable dataset to improve accuracy of the prediction and generalizability. By improving the accuracy of mortality risk predictions, the proposed model could provide a valuable tool to support medical professionals in making critical decisions and saving lives.

REFERENCES

[1] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," Neural Comput Appl, vol. 8, 2021, doi: 10.1007/s00521-020-05626-8.

[2] A. H. Elsheikh et al., "Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia," Process Safety and Environmental Protection, vol. 149, pp. 223–233, 2021, doi: 10.1016/j.psep.2020.10.048.

[3] H. S. Hota, R. Handa, and A. K. Shrivas, 27 - COVID-19 pandemic in India: forecasting using machine learning techniques. Elsevier Inc., 2021. doi: 10.1016/B978-0-12-824536-1.00030-7.

[4] X. Yang, "Nature-inspired optimization algorithms : Challenges and open problems," J Comput Sci, vol. 46, p. 101104, 2020, doi: 10.1016/j.jocs.2020.101104.

[5] H. Chao et al., "Integrative analysis for COVID-19 patient outcome prediction," Med Image Anal, vol. 67, p. 101844, 2021, doi: 10.1016/j.media.2020.101844.

[6] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," Inform Med Unlocked, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[7] M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," Chaos Solitons Fractals, vol. 139, p. 110050, 2020, doi: 10.1016/j.chaos.2020.110050.

[8] H. Burdick et al., "Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial," Comput Biol Med, vol. 124, p. 103949, 2020, doi: 10.1016/j.compbiomed.2020.103949.

[9] J. A. Guzmán-Torres, E. M. Alonso-Guzmán, F. J. Domínguez-Mota, and G. Tinoco-Guerrero, "Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that increase the risk of death using Machine learning, the case of Mexico," Results Phys, vol. 27, 2021, doi: 10.1016/j.rinp.2021.104483.

[10] A. Gumaei et al., "A Decision-Level Fusion Method for COVID-19 Patient Health Prediction," Big Data Research, vol. 1, p. 100287, 2021, doi: 10.1016/j.bdr.2021.100287.

[11] H. Al-Najjar and N. Al-Rousan, "A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea," Eur Rev Med Pharmacol Sci, vol. 24, no. 6, pp. 3400–3403, 2020, doi: 10.26355/eurrev_202003_20709.

[12] S. Aktar et al., "Machine learning approaches to identify patient comorbidities and symptoms that increased risk of mortality in covid-19," Diagnostics, vol. 11, no. 8, Aug. 2021, doi: 10.3390/diagnostics11081383.

[13] F. T. Fernandes, T. A. de Oliveira, C. E. Teixeira, A. F. de M. Batista, G. Dalla Costa, and A. D. P. Chiavegatto Filho, "A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil," Sci Rep, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-82885-y.

[14] H. Walia and S. Jeevaraj, "Early Mortality Risk Prediction in Covid-19 Patients Using an Ensemble of Machine Learning Models," 2021 International Conference on Computational Performance Evaluation, ComPE 2021, pp. 965–970, 2021, doi: 10.1109/ComPE53109.2021.9751945.

[15] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," Inform Med

Unlocked, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[16] Z. Ceylan, "Short-term prediction of COVID-19 spread using grey rolling model optimized by particle swarm optimization," Appl Soft Comput, vol. 109, p. 107592, 2021, doi: 10.1016/j.asoc.2021.107592.

