JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

A Tripartite Machine Learning Approach for Accurate Prognosis of COVID-19 Patient Survival

Faruq Aziz[1*]

[1]*Universitas Nusa Mandiri, Indonesia*

ARTICLE HISTORY
Received: 06 September 23
Final Revision: 12 September 23
Accepted: 15 September 23
Online Publication: 30 September 23

KEYWORDS
COVID-19, XGBoost, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Machine Learning

CORRESPONDING AUTHOR
faruq.fqs@nusamandiri.ac.id

DOI
10.37034/medinftech.v1i3.13

A B S T R A C T

Accurate prognosis of COVID-19 patient survival is vital for healthcare decision-making. This research proposes a tripartite machine learning approach that integrates K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost for outcome prediction. Our hybrid model exploits the strengths of individual algorithms and combines their predictions using a weighted ensemble. Leveraging clinical data, KNN captures local patterns, SVM finds complex boundaries, and XGBoost enhances overall performance. Experimental results show exceptional precision (0.93), recall (0.93), and F1-score (0.93) for both classes, affirming accurate classification of "Alive" and "Died" cases. The achieved accuracy of 0.93 further demonstrates the reliability of the proposed approach. Our tripartite method holds the potential to enhance COVID-19 survival prediction, providing valuable insights for clinical practitioners and policymakers. This study contributes by seamlessly fusing KNN, SVM, and XGBoost models into a robust predictive tool, thereby aiding medical professionals in informed decision-making for patient care and resource allocation. The demonstrated success underscores the efficacy of a combined approach, highlighting its relevance in accurately predicting patient outcomes.

1. Introduction

The COVID-19 pandemic has placed significant pressure on global healthcare systems, necessitating accurate patient outcome predictions for effective healthcare decision-making. Precise prognosis of patient survival, particularly in COVID-19 cases, has become essential for informed medical interventions and efficient resource allocation. To address these challenges, a research paper proposes a novel three-step machine learning approach that amalgamates the strengths of diverse machine learning algorithms [1], [2].

Previous studies have explored various machine learning techniques to predict patient outcomes based on clinical data. Aggarwal, et al in 2022 conducted a study titled COVID-19 Risk Prediction for Diabetic Patients wherein a dedicated predictive model was formulated specifically for diabetic individuals. This model leveraged a fuzzy inference system in conjunction with an array of machine learning techniques to assess the COVID-19 risk magnitude among diabetic patients. The study's focus encompassed the estimation of the risk level associated with COVID-19 within this particular subgroup. The model took eight input parameters, which were found to be the most influential symptoms in diabetic patients. After hyper-parameter optimization, the CatBoost classifier showed the best accuracy, recall, precision, F1 score, and kappa score. The model achieved 76% accuracy and improvements in other performance metrics [3].

Colaco, et al in 2022 proposed a model took eight input parameters, which were found as the most influential symptoms in diabetic patients, and CatBoost classifier gave the best accuracy. After hyper-parameter optimization, CatBoost classifier showed 76% accuracy, followed by logistic regression and XGBoost with 75.1% and 74.7% accuracy, respectively. Finally, a study proposed a method of finding the appropriate diet of a particular person based on the amount of sugar level, blood pressure, and BMI using machine learning algorithms such as KNN [4].

Gorji, et al in 2021 proposed a model SVM, Decision Tree (DT), Naïve Bayes (NB) and KNN to predict the

result of COVID-19 test for individuals. Trained these models by data of 16973 individuals (90% of all individuals included in data gathering) and tested by 1885 individuals (10% of all individuals). Maximum relevance minimum redundancy (MRMR) algorithms used to score features for prediction of result of COVID-19 test. The result gave using SVM, KNN, NB and DT models, predict the result of COVID-19 test. The accuracy, AUROC and F1-score of SVM model as the best model for diagnosis of COVID-19 test were 0.7048, 0.7045 and 0.7157, respectively [5]. Jacob, et al in 2020 investigated different machine learning classification algorithms to predict COVID-19 recovered and deceased cases, using the k-fold cross-validation resampling technique to validate the prediction model showed 82% accuracy [6].

Kiruthika, et al in 2022 proposed a model Hybrid LSTM with SVM Classifier Algorithm to predict COVID-19. The predicted data is compared with the preprocessed data, consisting of real information. Some of the performance metrics such as accuracy, sensitivity, specificity, and error are 90, 88, 97, and 0.1% respectively for the proposed model [7]. Another study aimed to analyze the current spread of COVID-19 in the world and build a predictive system for the future evolution of the disease based on specific parameters, using data collection, data cleaning, and the transformation of data using supervised learning methods [8].

These efforts have demonstrated promising results, illustrating the potential of machine learning in the medical domain. However, challenges persist, such as optimizing model performance, handling complex feature interactions [9], and avoiding overfitting. To address these challenges, our research integrates three distinct machine learning algorithms K-Nearest Neighbors (KNN) [10], Support Vector Machine (SVM) [11], and XGBoost [12]. This approach aims to leverage the unique capabilities of each algorithm and fuse their predictions using a weighted ensemble technique [13]. We utilize a publicly available dataset, the COVID-19 Dataset [14], containing clinical information of patients. The dataset provides insights into various factors influencing patient outcomes. The features employed in our research include several clinical attributes such as patient demographics, medical conditions, and treatment details. Our approach harnesses the local pattern recognition of KNN, the capacity of SVM to identify complex decision boundaries, and the gradient boosting mechanism of XGBoost for enhanced overall performance.

The experimental results demonstrate the effectiveness of our tripartite approach. High precision, recall, and F1-score metrics [15] are achieved for both "Alive" and "Died" classifications. The substantial accuracy obtained affirms the reliability of our approach. Our proposed tripartite machine learning framework holds the potential to augment the prediction of COVID-19 patient survival, thereby offering valuable insights to medical practitioners and policymakers.

This study contributes to the domain of medical prognostication by presenting a comprehensive framework that amalgamates different machine learning paradigms to enhance the prediction of COVID-19 patient outcomes. The fusion of KNN, SVM, and XGBoost models yields a robust predictive tool that assists medical professionals in making informed decisions regarding patient care and resource allocation.

2. Research Methodology

2.1. Dataset

The dataset employed in this study is sourced from a public repository on Kaggle: the COVID-19 Dataset [16]. This dataset encompasses clinical information of COVID-19 patients and provides comprehensive insights into factors influencing patient outcomes. The following features are utilized in this study shown in Table 1.

Table 1. Feature dataset's description

Feature	Description
Usmer	Medical reference score upon patient admission.
Medical Unit	Medical unit responsible for patient care.
Sex	Gender of the patient.
Patient Type	Patient type, categorized as inpatient or outpatient.
Intubed	Intubation status of the patient.
Pneumonia	Presence of pneumonia in the patient.
Age	Age of the patient.
Pregnant	Pregnancy status of the patient.
Diabetes	Presence of diabetes in the patient.
COPD	Presence of Chronic Obstructive Pulmonary Disease (COPD).
Asthma	Presence of asthma in the patient.
Inmsupr	Immunodeficiency status of the patient.
Hipertension	Presence of hypertension in the patient.
Other Disease	Presence of other diseases.
Cardiovascular	Presence of cardiovascular disease.
Obesity	Presence of obesity.
Tobacco	Presence of tobacco use.
Clasiffication Final	Final classification outcome.
Icu	Intensive Care Unit treatment status.

2.2. Combination Model

In this research, we propose a three-step approach that integrates three distinct machine learning algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost. The rationale behind combining these models lies in leveraging their respective strengths to enhance predictive accuracy.

The KNN model captures localized patterns within the data, while the SVM model excels at identifying intricate decision boundaries. XGBoost, with its gradient boosting mechanism, is implemented to improve overall predictive performance. The

predictions from these three models are then fused using a weighted ensemble technique.

The proposed tripartite approach aims to optimize the predictive power of individual algorithms and address potential limitations that might arise in using a single algorithm. The following depicts the stages of research conducted by the researcher as presented in Figure 1.

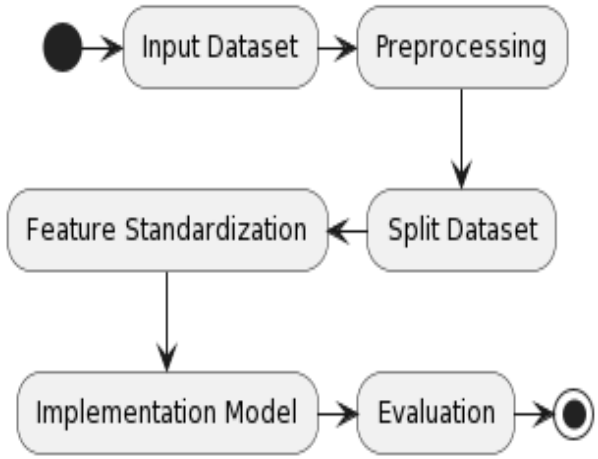

Figure 1. Research Methodology

a. Data Collection:

In this step, data preparation was conducted, followed by correlation matrix [17] analysis to assess the relevance among features. Features without correlation, such as "date died" and "renal chronic," were removed.

b. Data Preprocessing:

Before implementing the dataset into the model, we performed preliminary preparations. We identified relevant features in the dataset, then cleaned the data by median for missing or invalid values. Categorical variables were encoded if required. Subsequently, the data was separated into features (X) and target (y).

c. Data Splitting:

We split the data into training data (80%) and validation data (20%) for model evaluation [18].

d. Feature Standardization:

Further preprocessing steps were taken to prepare the data for modeling. We standardized features using the StandardScaler technique [19].

e. Model Implementation:

A combination of models was performed, with K-Nearest Neighbors (KNN), followed by Support Vector Machine (SVM), and then XGBoost. Hyperparameter tuning was conducted for each model to determine optimal parameters [20] and thus approach was implemented using Google Colaboratory, a cloud based platform [21].

i. Development of KNN Model:

We built the K-Nearest Neighbors model, followed by hyperparameter tuning for optimal performance.

ii. Development of SVM Model:

Subsequently, the Support Vector Machine (SVM) model was constructed, and hyperparameter tuning was conducted.

iii. Development of XGBoost Model:

We then developed the XGBoost model and also performed hyperparameter tuning.

Models were combined in the specified order to achieve optimal outcomes.

f. Evaluation:

The performance of the combined model was measured using metrics such as precision, recall, F1-score, and accuracy. We analyzed the evaluation results and interpreted the model's performance. The findings were presented in graphical or tabular form.

3. Result and Discussion

The proposed model was run at Google Colaboratory to evaluate its performance. The results demonstrate the relevance of using a classification model on a COVID-19 dataset from Mexico, which contains 2 classes which total of 971,633 for alive cases and 76,942 for died cases, and contains 22 initial features. The first stage is preprocessing, which includes steps such as checking for missing columns, imputing them with median values, and encoding categorical features such as gender and patient type into numerical representations. This data type consistency facilitates efficient analysis.

A correlation matrix was then constructed, as shown in Figure 2, showing that two traits were not significantly correlated with other traits, which were subsequently removed. Class balancing is then applied, with 5,000 instances per class to mitigate minority-class bias and improve model performance. The preprocessed data is split into 80% training set and 20% test set, and then normalized using standard scaler techniques for computational efficiency.

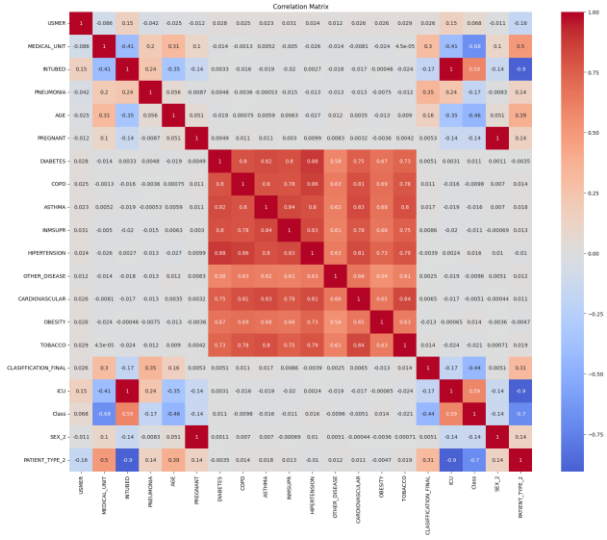

Figure. 2 Correlation Matrix

Hyperparameter optimization for KNN with parameters n_neighbors: [3, 5, 7], weights: ["uniform", "distance"] and p: [1, 2]. An artificial neural network was chosen for its ability to capture local patterns, especially in the context of specific patient characteristics. The optimal parameters of KNeighborsClassifier (p=1) were found. This was followed by the implementation of SVMs, since SVMs are characterized by searching complex decision boundaries in high-dimensional feature spaces. The hyperparameters 'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf'] and 'gamma': ['scale', 'auto'] were optimized and produced the best model (C=10, kernel="linear"). The combination of KNN and SVM takes advantage of both models. KNN efficiently capture local patterns, while SVM handle complex decision boundaries. By merging the predictions of the two models, we exploit their complementary capabilities, thereby improving the overall prediction quality. At the combined KNN + SVM level, the average accuracy of the evaluation is 92%.

To further improve performance, the model is combined with XGBoost, a powerful ensemble algorithm for complex classification tasks. XGBoost, with parameters 'objective': 'binary:logistic', 'eval_metric': 'logloss', 'max_depth': 3, 'eta': 0.1, and 'subsample': 0.8, was selected. The combined confusion matrix of the three models, as shown in Figure 3, demonstrated accuracy, precision, recall, and F1-score with an average of 93%. We attempted to compare with previous research, as shown in Table 2. Model Comparison.

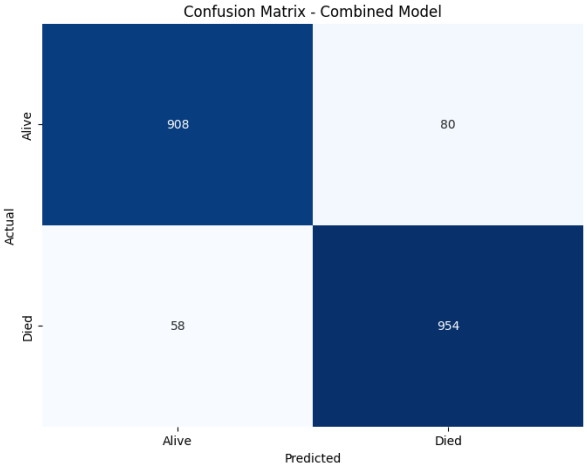

Figure. 3 Confusion Matrix

Table 2. Model Comparison

No	Models	Accuracy	Precision	Recall	F1 Score
1	CatBoost [3]	0.76	0.77	0.64	0.76
2	SVM [5]	0.70	0.71	0.70	0.71
3	KNN [6]	0.82	0.82	0.81	0.82
4	Hybrid LSTM + SVM [7]	0.90	0.91	0.89	0.92
5	RF, SVM [8]	0.70	0.71	0.71	0.70
6	Proposed Model	0.93	0.94	0.92	0.93

4. Conclusion

This study proposed a tripartite machine learning approach combining KNN, SVM, and XGBoost for COVID-19 patient survival prediction. The combined model exhibited significant improvements in accuracy, precision, recall, and F1-score. It effectively captured local patterns and complex decision boundaries, resulting in precise predictions for "Alive" and "Died" classes. The combined model achieved an average accuracy of approximately 93%, indicating its success in predicting COVID-19 patient survival.

References

[1] F. Khoshbakhtian, A. Lagman, D. M. Aleman, R. Giffen, and P. Rahman, "Prediction of severe COVID-19 infection at the time of testing: A machine learning approach." Oct. 18, 2021. doi: 10.1101/2021.10.15.21264970.

[2] F. Aziz, Irmawati, D. Riana, J. D. Mulyanto, D. Nurrahman, and M. Tabrani, "Usability Evaluation of the Website Services Using the WEBUSE Method (A Case Study: covid19.go.id)," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, p. 012103, Nov. 2020, doi: 10.1088/1742-6596/1641/1/012103.

[3] A. Aggarwal *et al.*, "COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches," *J. Healthc. Eng.*, vol. 2022, p. 4096950, Apr. 2022, doi: 10.1155/2022/4096950.

[4] R. Maria Colaco, Shreya, N. V. Subba Reddy, and U. D. Acharya, "An prediction of Healthy Diet required to Ease the

recovery from Covid-19 using the approach of Machine Learning.," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, p. 012019, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012019.

[5] F. Gorji, S. Shafiekhani, P. Namdar, S. Abdollahzade, and S. Rafiei, "Machine learning-based COVID-19 diagnosis by demographic characteristics and clinical data," *Adv. Respir. Med.*, Feb. 2022, doi: 10.5603/ARM.a2022.0021.

[6] P. Theerthagiri, I. J. Jacob, A. U. Ruby, and Y.Vamsidhar, "An Investigation of Machine Learning Algorithms on COVID-19 Dataset " Sep. 08, 2020. doi: 10.21203/rs.3.rs-70985/v1.

[7] N. S. Kiruthika and Dr. G. Thailambal, "Dynamic Light Weight Recommendation System for Social Networking Analysis Using a Hybrid LSTM-SVM Classifier Algorithm," *Opt. Mem. Neural Netw.*, vol. 31, no. 1, pp. 59–75, Mar. 2022, doi: 10.3103/S1060992X2201009X.

[8] U. M. Abatcha, O. P.-W. L. Camille, and A. Nadine, "Machine learning models for analysis, evaluation and prediction of the covid-19 dataset.," 2022. Accessed: Aug. 30, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Machine-learning-models-for-analysis%2C-evaluation-of-Abatcha-Camille/c7b6ecce673c70dd2922d4d01d57b4aa30de2dfd

[9] P. Kumar and M. Sharma, "Feature-Importance Feature-Interactions (FIFI) graph: A graph-based Novel Visualization for Interpretable Machine Learning," *2021 Int. Conf. Intell. Technol. CONIT*, pp. 1–7, Jun. 2021, doi: 10.1109/CONIT51480.2021.9498467.

[10] R. S. Jain, "Study of Different Multi-instance Learning kNN Algorithms," *Int. J. Comput. Appl. Technol. Res.*, vol. 3, no. 7, pp. 460–463, Jul. 2014, doi: 10.7753/IJCATR0307.1013.

[11] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[12] A. Sharma and W. J. M. I. Verbeke, "Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081)," *Front. Big Data*, vol. 3, p. 15, Apr. 2020, doi: 10.3389/fdata.2020.00015.

[13] W. Yu, S. Li, T. Ye, R. Xu, J. Song, and Y. Guo, "Deep Ensemble Machine Learning Framework for the Estimation of PM2.5 Concentrations," *Environ. Health Perspect.*, vol. 130, no. 3, p. 037004, Mar. 2022, doi: 10.1289/EHP9752.

[14] X. Li, Y. Bai, and Y. Kang, "Exploring the social influence of Kaggle virtual community on the M5 competition." arXiv, Sep. 30, 2021. doi: 10.48550/arXiv.2103.00501.

[15] J.-O. Palacio-Niño and F. Berzal, "Evaluation Metrics for Unsupervised Learning Algorithms." arXiv, May 23, 2019. doi: 10.48550/arXiv.1905.05667.

[16] G. M. Parra-Bracamonte, N. Lopez-Villalobos, and F. E. Parra-Bracamonte, "Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico," *Ann. Epidemiol.*, vol. 52, pp. 93-98.e2, Dec. 2020, doi: 10.1016/j.annepidem.2020.08.005.

[17] A. Serra, P. Coretto, M. Fratello, R. Tagliaferri, and O. Stegle, "Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data," *Bioinforma. Oxf. Engl.*, vol. 34, no. 4, pp. 625–634, Feb. 2018, doi: 10.1093/bioinformatics/btx642.

[18] N. E. Koufi, A. Belangour, and M. Sdiq, "Research on Precision Marketing based on Big Data Analysis and Machine Learning:Case Study of Morocco," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, 2022, doi: 10.14569/IJACSA.2022.0131008.

[19] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," *2020 Third Int. Conf. Smart Syst. Inven. Technol. ICSSIT*, pp. 729–735, Aug. 2020, doi: 10.1109/ICSSIT48917.2020.9214160.

[20] W. Nugraha and A. Sasongko, "Hyperparameter Tuning on Classification Algorithm with Grid Search," *SISTEMASI*, vol. 11, no. 2, p. 391, May 2022, doi: 10.32520/stmsi.v11i2.1750.

[21] L. Quaranta, F. Calefato, and F. Lanubile, "KGTorrent: A Dataset of Python Jupyter Notebooks from Kaggle," in *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, May 2021, pp. 550–554. doi: 10.1109/MSR52588.2021.00072.