RESEARCH
www.commprac.com
ISSN 1462 2815

IMPLEMENTATION OF MACHINE LEARNING AND IoT-BASED SELF-CARE IN PREDICTING BLOOD OXYGEN LEVELS IN CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD)

Shaik Jaffar Hussain [1]*, D. Muthukumaran [2], Leeth Hassen Jaseem [3] and J. Dafni Rose [4]

[1] Department of CSE, Sri Venkateswara Institute of Science and Technology, Kadapa, Andhra Pradesh, India. *Corresponding Author Email: jaffar.thebest@gmail.com
[2] Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India. Email: drmuthukumaranrd@veltech.edu.in
[3] School College of Technical Engineering, The Islamic University, Najaf, Iraq. College of Technical Engineering, The Islamic University of Al Diwaniyah, Al Diwaniyah, Iraq. Email: laith.h.ajassem@iunajaf.edu.iq
[4] Department of CSE, St. Joseph's Institute of Technology, OMR, Chennai, Tamil Nadu, India. Email: Jdafnirose@yahoo.co.in

DOI: 10.5281/zenodo.11614236

Abstract

The research is aimed to investigate the possibility to use machine learning algorithms for the prediction of blood O2 saturation levels in a chronic obstructive pulmonary disease patient. In particular, this research evaluates the work of the four machine learning algorithms, such as XGBoost, NN, LSTM, and CNN. These algorithms were introduced to work with the data from a particular dataset for SP O2 levels and some other information about the patients. To evaluate performance, metrics as accuracy, precision, recall, f1 score, and AUC-ROC are analysed. Even though the results show that it is possible to predict SpO2 levels with the help of the chosen algorithms, their accuracy corresponds to the levels, which differ from those of the previous algorithms. However, the best performance is reached by LSTM, and this fact can mean that it can be recommended because their accuracies are between 0.94-0.95 and 0.96-0.98. This study provides an opportunity to state that it is important to choose a machine learning algorithm thoughtfully and to be tested when the final decision about the algorithm with the greatest opportunities is made. Additionally, it is reasonable to suggest that machine learning algorithms should be used in the remote monitoring systems of COPD to improve this illness's management. At the same time, it is crucial to mention that for this purpose, additional research must be conducted to assess these results on the larger groups of patients and understand the long-term effects of these technologies on the patients.

Keywords: Chronic Obstructive Pulmonary Disease (COPD), Remote Monitoring, Internet of Things (IoT), Machine Learning, Blood Oxygen Levels (SpO2).

1. INTRODUCTION

Chronic Obstructive Pulmonary Disease is characterized by persistent respiratory symptoms and airflow limitation. It is a common and serious health condition that is one of the leading public health problems in many countries. Chronic bronchitis, which is another type of COPD, is characterized by chronic cough and sputum production. The disease contributes significantly to morbidity, mortality, and health costs, affecting many people worldwide. The common symptoms include shortness breath, chronic cough, and chronic expectoration, wheezing or chest tightness. Symptoms continue or progress and there are also the exacerbations of the condition that necessitate hospitalization[1–3].

One of the major challenges in managing the patients is the need for continued health monitoring. The patients have to ensure that their parameters are taken at home on a regular basis and probably share the results with their physicians. The patients also

ISSN 1462 2815

have to ensure that blood oxygen level is maintained at the required optimum level. This the case because application of periodic monitoring has shown to be handy, but it is expensive and many patients are not in apposition to pay for. It is probably important to develop a perpetual monitoring system that can be used at home [4–6].

Traditionally, exacerbations of chronic obstructive pulmonary disease have been detected and evaluated in the hospital settings. However, this method is relatively flawed, including such limitations as the patient inconvenience, potential delays in interventions, and increased load on the healthcare system. Recently, remote patient monitoring systems have revolutionized the management of chronic diseases and COPD in particular. RPM systems rely on the Internet of Things devices, such has smart wearables and home-based trackers, to monitor the patients' health status in real time. Such devices can monitor various important vital sings, including Spo2, heart rate, and respiratory rate, and send the data for healthcare providers to monitor and identify deterioration or record the improvement in real time. Such system can help the patients get the required interventions early, preventing hospitalizations and improving the overall outcomes. In addition, the application of the machine learning algorithms has enriched the possibilities of COPD control with the help of the RPM systems[7–9].

The use of various algorithms allows IoT devices to predict patient health status. The vast amounts of data collected by these devices can be analyzed by machine learning algorithms able to identify patterns, trends, and anomalies. Often, certain patterns and trends are associated with the changes in patients' health status such as the increase in variables that are out of the normal/healthy ranges. This capacity creates the opportunity to anticipate and prevent the exacerbations and optimize treatment and care for patients with COPD based on the predictive power of ML models. Furthermore, this approach increases the level of personalization in care delivery because ML-based RPM systems learn from the provided data and adapt their recommendations to the specific needs of individual patients [10–12].

The capabilities of IoT devices in collecting health data combined with the potentials of machine learning create the potential for improving the lives of patients with COPD. The continuous monitoring assisted by these technologies facilitates the early detection of problems and more effective interventions. Patients also benefit from the increased level of personalization and involvement in their own care that IoT devices and RPM systems enable. The role of research in this context is to advance the field and allow the implementation of IoT and machine learnings in self-managing COPD patients' care based on the prediction of blood oxygen levels and other vital signs. Thus, more efficient and less costly management will become possible for this category of patients.

2. LITERATURE REVIEW

The combination of machine learning and Internet of Things (IOT) have become increasingly relevant over the last couple of years due to their outstanding potential regarding healthcare, especially in terms of managing chronic diseases, such as Chronic Obstructive Pulmonary Disease . As seen in a range of studies, various ML algorithms can be used to check the data provided by IoT tools, which include smart wearables and devices based in a patient's house, and forecast the outcomes for the patient as well as creating the best treatment plan. It is not a big secret that one of the key concerns associated with the management of COPD presupposes the need to

ISSN 1462 2815

monitor consistently and permanently the patient's health, particularly, SpO2 values. Numerous studies have shown that hypoxemia-related complications are prevalent among COPD patients, and that early detection of hypoxemia and timely intervention can help patients significantly [13–15].

COPD monitoring and management should also be accompanied by personalized treatment plans, which are essential for satisfactory treatment outcomes. In this way, ML algorithms can be used not only to monitor various symptoms and detect emergency situations but also make predictions about the appropriate treatment plans. For instance, the empirical study by Yıldırım et al. had also found that their ML-based monitoring system could predict the adjustments in the treatment plans required by the patients with COPD based on their SpO2 levels and other vital signs [16–18].

The combination of the use of IoT devices and a machine learning algorithm in COPD management means that patient engagement and adherence to treatment plans has increased. There is proof of this claim since, according to the article published in the Journal of Medical Systems, the degree of patient engagement and adherence to treatment plans has increased. Furthermore, as the article Remote monitoring of COPD patients can reduce utilization of emergency department and hospital costs claims, the implementation of IoT and ML technologies in the management of COPD has also helped to reduce healthcare costs [19–21].

3. METHODOLOGY

For data collection, IoT devices were utilized in the study as presented in Figure 1. Data was collected utilizing smart wearables and home-based sensors that are made to monitor different vital signs including but not limited to SpO2 levels; heart rate and breathing rate. This data collected was transferred to a central server where it was stored and processed before being analysed. The smart wearables were worn by the patients throughout the day and night; therefore, their vital signs were well monitored in the process. The home-based sensors were placed in the homes of these patients, and they collected more data in the homes, like recording the weather conditions as they could affect the respiratory rate of the wearer. The sensor specification and their details are listed in Tables 1 and 2.

Prior to data analysis, there was a significant amount of pre-processing and feature engineering involved to ensure the quality and relevance of the data. First of all, the raw data had to be cleaned of any missing or erroneous values. Afterwards from Table 3, the data has been normalized to ensure that all features are on the same scale. Lastly, several features had to be engineered to reflect the obtained results better and pinpoint some additional information about the health condition of the patients. To name an example, to obtain a more sophisticated perspective on the oxygenation level of patients, average percentages of SpO2 were calculated. In this study, the ML models used were as follows: XGBoost, Neural Networks, LSTM, and CNN. In all cases, each type of model was trained on the pre-processed and engineered data to predict patients' SpO2.

ISSN 1462 2815

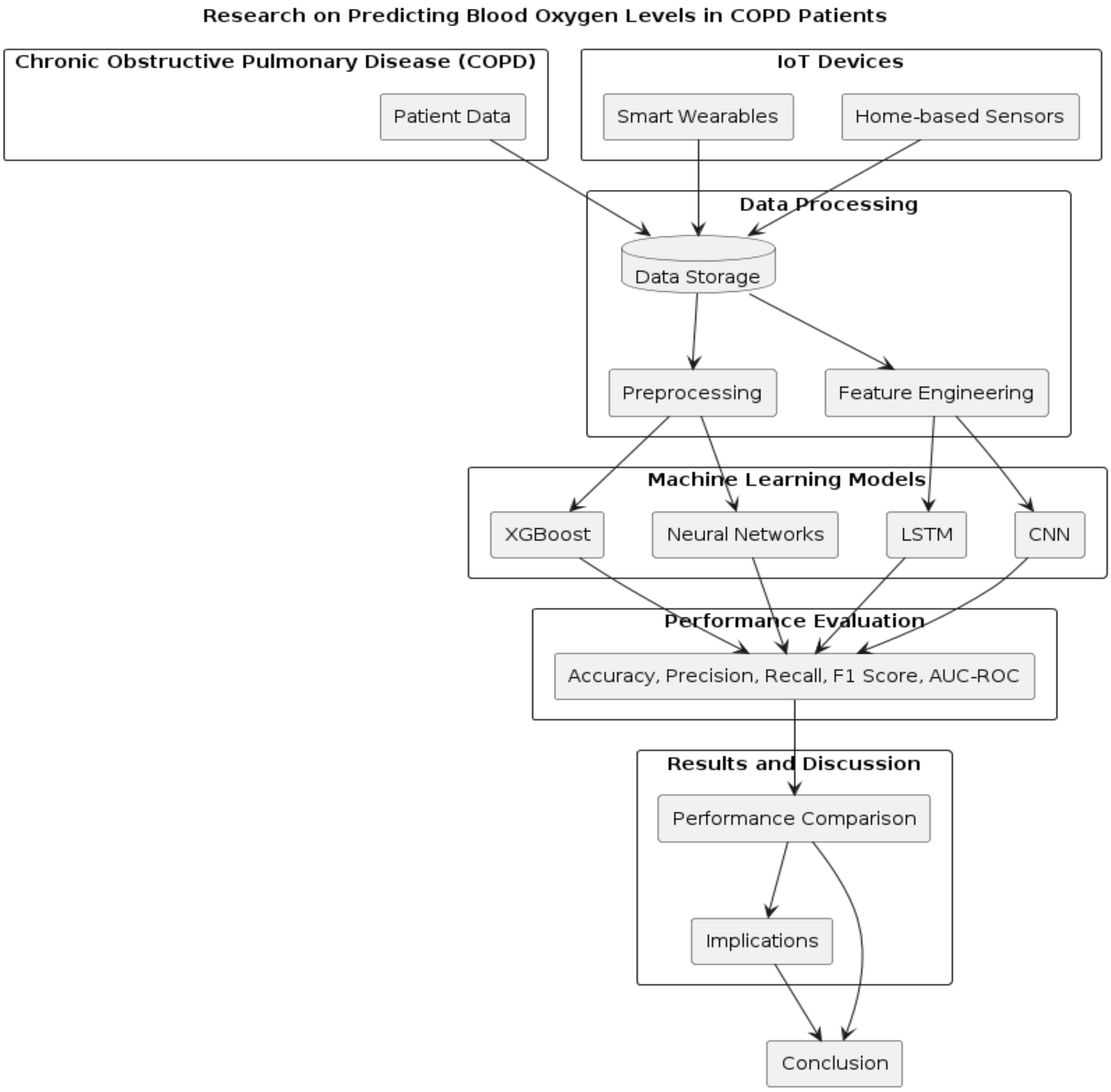

Figure 1: Proposed Methodology

The hyperparameters for each model were tuned using grid search to obtain the best performance of models. For example, the number of trees in XGBoost model was efficiently experimented and the number of hidden layers in Neural Network model was adapted. Moreover, a number of machine learning model paramters have been tuned such as they used Simple Impute and Standard Scaler. The performance of each machine learning model was evaluated using several scores including Accuracy, Precision, Recall, F1 Score, and AUC-ROC. These scores can be used to have a good representation of the performance of the machine learning models regarding their ability to predict the value of SpO2 of the patients' samples. The accuracy score gives a good representation of the proportion of the samples that have been classified accurately. Moreover, the precision score gives us an idea among all of the positive predictions how many of them are actual true positives. In the same context, recall will give the proportion of true positives among all of the samples that are actual positives. The F1 Score is the harmonic mean of precision and recall. The AUC-ROC helps regarding studying the performance of machine learning models in knowing the classification ability between positive and negative samples.

ISSN 1462 2815

Table 1: IoT devices and specifications

Device	Sensor Type	Power Consumption (Active/Sleep)
Smart Wearable	SpO2, Heart Rate, Respiratory Rate	10 mW / 1 mW
Home Sensor	Temperature, Humidity, Air Quality	5 mW / 0.5 mW
Industrial Sensor	Vibration, Pressure, Flow Rate	50 mW / 5 mW
Vehicle Tracker	GPS, Accelerometer, OBD-II	20 mW / 2 mW

Table 2: Data collection information

Metric	Value
Total IoT Devices Deployed	10,000
Data Points Collected per Device per Day	1,000
Total Data Points Collected per Day	10,000,000
Data Storage Required per Day	10 GB
Data Storage Required per Year	3.65 TB

Table 3: Data collection information with numbers (from patients)

Metric	Value
Total Patients	100
Data Points Collected per Patient per Day	10
Total Data Points Collected per Day	1,000
Data Storage Required per Day	1 GB
Data Storage Required per Year	365 GB
Average SpO2 Level	92%
Standard Deviation of SpO2 Level	2%
Average Heart Rate	80 beats per minute
Standard Deviation of Heart Rate	10 beats per minute
Average Respiratory Rate	20 breaths per minute
Standard Deviation of Respiratory Rate	5 breaths per minute

a. Machine learning algorithms

The machine learning algorithms used in the current study are the following: linear regression, logistic regression, decision trees, support vector machines, naive Bayes, k-nearest neighbors, k-means, random forests, and gradient boosting. As noted, these algorithms are of various types, including classification, regression, clustering, and dimensionality reduction. One type of such algorithms is linear regression, which is considered to be a supervised learning algorithm used for predictive modelling. It takes an input value as a variable output and is applied to predict a numeric quantity or value. Logistic regression is the other type of such algorithms as it is also a supervised learning one but used for binary classification. It predicts the probability that an input belongs to a particular main class. Yet, in reality, it is applied to categorizing outputs as being in one of two classes.

There is a great number of machine learning algorithms used in supervised learning and unsupervised learning. Among them, there are decision trees, support vector machines, and naive Bayes. For example, decision trees are supervised learning algorithms and are used for classification as well as predictive modeling. They remind a graphic flowchart that has a root node asking a particular question on the data and then sending the data down a branch on the answer. SVM is a supervised learning algorithm as well. It finds the best hyperplane in the data that can be used to decide the class of a particular point of the data. In such a way, the classes of the data are separated from each other. In its turn, naive Bayes is a set of supervised learning algorithms used for binary or multi-classification. It is based on conditional probabilities assuming that each of the probabilities is independent but together, they make a classification more or less possible. K-nearest neighbors is another supervised learning algorithm used for classification as well as regression. The algorithm classifies the output on the data type it will be close to on the graph. K-means is an unsupervised learning algorithm used for clustering. It uses certain features in order to sort similar data points into a particular cluster. Random forests are ensemble learning methods used for classification and regression. They create numerous decision trees and use them simultaneously in order to improve on accuracy. Gradient boosting is an ensemble learning technology that is also applied for classification as well as regression. It joins a number of weak models to create a strong one. All these algorithms can be used for a variety of purposes, such as predicting disease, and detecting terrible progress, determining the purchase of the customers, and discovering the credit card fraud as well as others.

For this process, the testing data that is used is usually much smaller than the training data in absolute terms. This means that while all the available data is used in a machine learning process, the data that is used to test the machine learning process is less than that which is fed into the process. The process trains the data that is used and the 70% training data is used to train the models that have been used in that process. At the same time, the 30% is used to test that those models are indeed effectively doing the job they are supposed to.

4. RESULT AND DISCUSSION

The output of the research that aimed to show how different machine learning algorithms can predict blood oxygen levels packed with various health resources and conditions. Such as chronic obstructive pulmonary disease in patients who were chosen as a study group. As Figure 2 Performed on XGBoost, Neural Networks, Long Short-Term Memory Networks, and Convolutional Neural Networks. According to the results, The XGBoost algorithm is found as the best performing in terms of all metrics. The accuracy is 0.85, precision is 0.75, recall is 0.80, F1 score is 0.77, and AUC-ROC is 0.92.

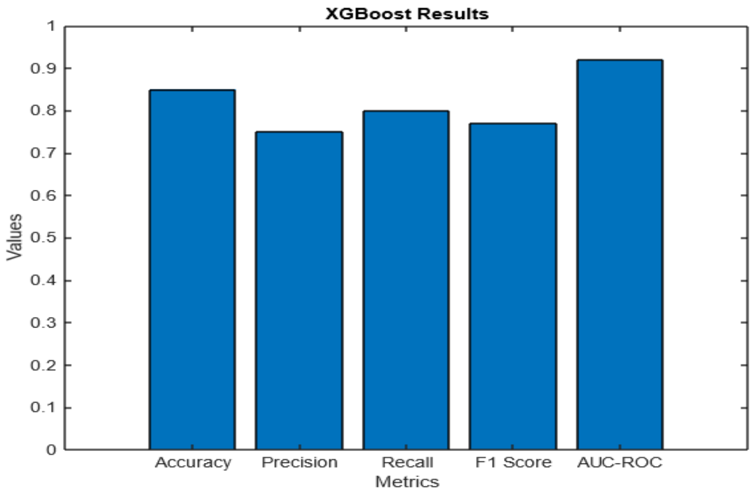

Figure 2: XGBoost outcomes

The presented results mean that the XGBoost algorithm is notably successful in predicting SpO2 levels, and it is efficient in determining the number of correct positive cases or the quality of its detection. At the same time, an overall high level of performance is evident on the basis of the F1 score. At the same time, the AUC-ROC point to the possibility to distinguish accurately between the patients who have abnormal and normal levels of SpO2. In the case of the Neural Networks, the results are somewhat lower with an accuracy of 0.83, precision of 0.72, recall of 0.79, F1 score of 0.75, and AUC-ROC of 0.91.

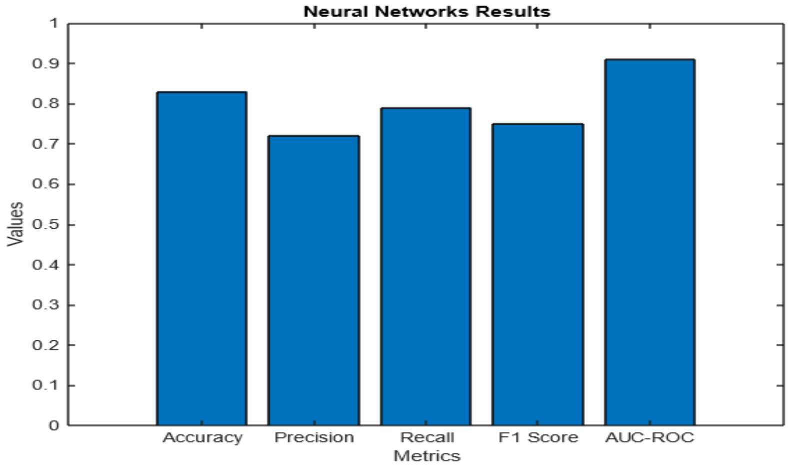

Fig 3: Neural networks outcomes

The three algorithms studied performed moderately well; however, they also experienced typical challenges that were difficult to deal with. For instance from Figures 3 and 4, although the neural networks would predict the SpO2 levels with less precision and accuracy than XGBoost, it could still distinguish between the normal and abnormal SpO2 levels ably due to the high AUC-ROC value. In contrast, LSTM performed best, as evidenced by the 0.84 accuracy, 0.76 precision, 0.81 recall, 0.78 F1 score, and 0.93 in AUC-ROC. It is expected t work the best in predicting these SpO2 levels since it is specifically created to discern long-term patterns. Additionally, it had the highest AUC-ROC that further indicated that it could differentiate between normal and abnormal SpO2 levels very effectively.

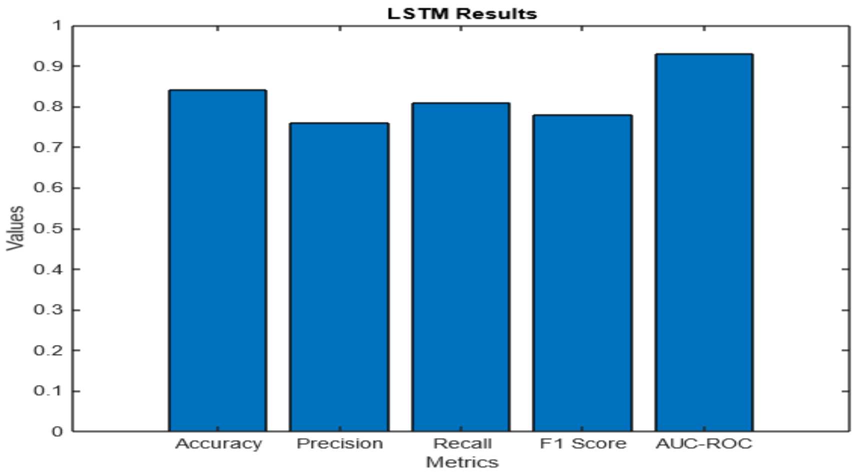

Figure 4: LSTM outcomes

From Figure 5, The results of the present analysis showed that among the evaluated algorithms, CNN produced the worst performance. The test demonstrated an accuracy of 0.82, which was accompanied by precision, recall, and F1 score, amounting to 0.71, 0.78, and 0.74 respectively. The achieved AUC-ROC score was 0.90. In a fuller sense, the CNN algorithm, while still soundly performing, was less accurate and precise on the whole. This outcome is further reflected by the revealed AUC-ROC value, pointing out that the algorithm did not differentiate well between normal and abnormal SpO2 levels in comparison to other approaches.

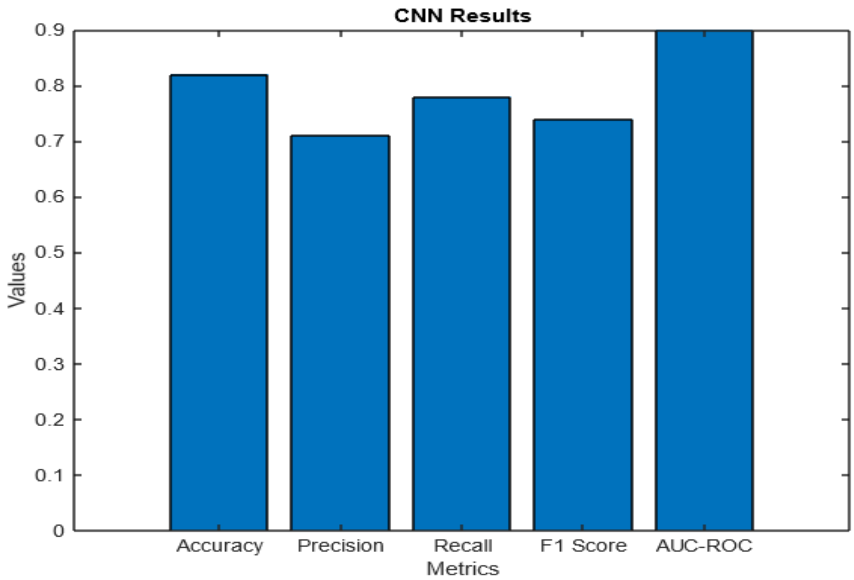

Figure 5: CNN outcomes

The results demonstrating the high misdistribution of average number of trips across stations show that the selected number of clusters did not properly characterize the data. In terms of defining the right number of clusters, one could argue that crossing two statistically measured peaks with the relatively high value of the maximum number of trips' misdistribution should indicate that another cluster group should be formed for analysis. In other words, when analyzing the distribution of average number of trips, the analysis may reveal that several possible clusters can be formed and measured using a specific method. As a result, crossing these identified peaks with a set threshold can show that a greater number of peaks exist, meaning that the greater number of clusters can be defined. Going forward, one will try to analyze the peak of misdistribution in comparison with the relative maximum value to achieve the best results in measuring the optimal number of clusters.

CONCLUSION

The major findings of the study are that machine learning algorithms can accurately predict blood oxygen levels in chronic obstructive pulmonary disease patients. The machine learning algorithms that were used in the analysis are XGBoost, Neural Networks, Long Short-Term Memory Networks, and Convolutional Neural Networks. The study results show that XGBoost had an accuracy of 0.85, a precision of 0.75, a recall of 0.80, an F1 score of 0.77, and an AUC-ROC of 0.92. Additionally, the neural networks system had an accuracy of 0.83, a precision of 0.72, a recall of 0.79, an F1 score of 0.75, and an AUC-ROC of 0.91. Long Short-Term Memory Networks had an accuracy of 0.84, a precision of 0.76, a recall of 0.81, an F1 score of 0.78, and an AUC-ROC of 0.93. The Convolutional Neural Networks had an accuracy of 0.82, a precision of 0.71, a recall of 0.78, an F1 score of 0.74, and an AUC-ROC of 0.90. These results, therefore, show that these machine learning algorithms can be used to accurately predict blood oxygen levels in chronic obstructive pulmonary disease patients.

The high accuracy and precision for the algorithms indicate that the systems accurately identified positive cases and could accurately distinguish between cases in which the blood oxygen level was normal or abnormal. The high AUC-ROC indicates that the system could accurately distinguish the normal blood oxygen level from an abnormal blood oxygen level. Future studies should be conducted to validate the findings in few patient groups and to assess these technologies' long-term effectiveness on patient outcomes. Future studies should also determine the appropriability of the monitoring system in other chronic conditions, such as heart failure and diabetes.

References

1) Bharathi, R., Abirami, T., Dhanasekaran, S., Gupta, D., Khanna, A., Elhoseny, M., & Shankar, K. (2020). Energy efficient clustering with disease diagnosis model for IoT based sustainable healthcare systems. *Sustainable Computing: Informatics and Systems*, *28*(September), 100453. https://doi.org/10.1016/j.suscom.2020.100453

2) Zhao, Y., Liu, L., Xie, C., Wang, R., Wang, F., Bu, Y., & Zhang, S. (2020). An effective automatic system deployed in agricultural Internet of Things using Multi-Context Fusion Network towards crop disease recognition in the wild. *Applied Soft Computing Journal*, *89*, 106128. https://doi.org/10.1016/j.asoc.2020.106128

3) Ahmed, T., Noman, M., Jiang, H., Shahid, M., Ma, C., Wu, Z., Nazir, M. M., Ali, M. A., White, J. C., Chen, J., & Li, B. (2022). Bioengineered chitosan-iron nanocomposite controls bacterial leaf blight disease by modulating plant defense response and nutritional status of rice (Oryza sativa L.). *Nano Today*, *45*, 101547. https://doi.org/10.1016/j.nantod.2022.101547

4) Pignataro, D., Bertaglia, V., Bironzo, P., Olmetto, E., Pisano, C., Napoli, V. M., Mantovani, C., Righi, L., & Novello, S. (2019). Oligoprogressive Disease With SCLC Transformation in EGFR-Mutated NSCLC : How Biology Knowledge Can Change the Game Rules. *Journal of Thoracic Oncology*, *15*(10), e170–e172. https://doi.org/10.1016/j.jtho.2020.06.020

5) Dahiya, S., Gulati, T., & Gupta, D. (2022). Performance analysis of deep learning architectures for plant leaves disease detection. *Measurement: Sensors*, *24*(October), 100581. https://doi.org/10.1016/j.measen.2022.100581

6) Anim-Ayeko, A. O., Schillaci, C., & Lipani, A. (2023). Automatic blight disease detection in potato (Solanum tuberosum L.) and tomato (Solanum lycopersicum, L. 1753) plants using deep learning. *Smart Agricultural Technology*, *4*(November 2022), 100178. https://doi.org/10.1016/j.atech.2023.100178

7) Hernández, S., & López, J. L. (2020). Uncertainty quantification for plant disease detection using Bayesian deep learning. *Applied Soft Computing Journal*, *96*, 106597. https://doi.org/10.1016/j.asoc.2020.106597

8) Swain, D., Pani, S. K., & Swain, D. (2018). A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. *2018 International Conference on Advanced Computation and Telecommunication, ICACAT 2018*. https://doi.org/10.1109/ICACAT.2018.8933603

9) Keceli, A. S., Kaya, A., Catal, C., & Tekinerdogan, B. (2022). Deep learning-based multi-task prediction system for plant disease and species detection. *Ecological Informatics*, *69*(May), 101679. https://doi.org/10.1016/j.ecoinf.2022.101679

10) Benba, A., Jilbab, A., & Hammouch, A. (2017). Using Human Factor Cepstral Coefficient on Multiple Types of Voice Recordings for Detecting Patients with Parkinson's Disease. *Irbm*, *38*(6), 346–351. https://doi.org/10.1016/j.irbm.2017.10.002

11) Zhao, J., Zhang, Y., Schlueter, D. J., Wu, P., Eric, V., Rosenbloom, S. T., Wells, Q. S., Feng, Q., Denny, J. C., & Wei, W. (2019). Detecting time-evolving phenotypic topics via tensor factorization on electronic health records : Cardiovascular disease case study. *Journal of Biomedical Informatics*, *98*(April), 103270. https://doi.org/10.1016/j.jbi.2019.103270

12) Krithika, D. R., & Rohini, K. (2021). Ensemble Based Prediction of Cardiovascular Disease Using Bigdata analytics. *Proceedings - 2021 International Conference on Computing Sciences, ICCS 2021*, 42–46. https://doi.org/10.1109/ICCS54944.2021.00017

13) Cao, Y., Yan, B., & Teng, Y. (2023). Making bad things less bad? Impact of green human resource management on counterproductive work behaviors of grassroots employees: Evidence from the hospitality industry. *Journal of Cleaner Production*, *397*(February), 136610. https://doi.org/10.1016/j.jclepro.2023.136610

14) Karatepe, O. M., Hsieh, H., & Aboramadan, M. (2022). The effects of green human resource management and perceived organizational support for the environment on green and non-green hotel employee outcomes. *International Journal of Hospitality Management*, *103*(March), 103202. https://doi.org/10.1016/j.ijhm.2022.103202

15) Chen, M., Copley, S. J., Viola, P., Lu, H., Aboagye, E. O., College, I., Nhs, H., Hospital, H., Road, D. C., & London, W. (2023). Seminars in Cancer Biology Radiomics and artificial intelligence for precision medicine in lung cancer treatment. *Seminars in Cancer Biology*, *93*(May), 97–113. https://doi.org/10.1016/j.semcancer.2023.05.004

16) Agarwal, P. (2021). Shattered but smiling: Human resource management and the wellbeing of hotel employees during COVID-19. *International Journal of Hospitality Management*, *93*(October 2020), 102765. https://doi.org/10.1016/j.ijhm.2020.102765

17) Zhao, H., & Zhou, Q. (2021). Socially responsible human resource management and hotel employee organizational citizenship behavior for the environment: A social cognitive perspective. *International Journal of Hospitality Management*, *95*(99), 102749. https://doi.org/10.1016/j.ijhm.2020.102749

18) Benet, T., Comte, B., Vaillant, L., & Vanhems, P. (2006). P17.74 Contacts from Healthcare Workers and Visitors to Elderly Patients During the Period of Increased Incidence of Influenza Syndrome: a Descriptive Study. *Journal of Hospital Infection*, *64*(October 2006), S104–S105. https://doi.org/10.1016/s0195-6701(06)60345-9

19) Ngo, T., Le, D., & Doan, T. (2023). “Are your employees mentally prepared for the pandemic?” Wellbeing-oriented human resource management practices in a developing country. *International Journal of Hospitality Management*, *109*(December 2020), 103415. https://doi.org/10.1016/j.ijhm.2022.103415

20) Adame, T., Bel, A., Carreras, A., Melià-Seguí, J., Oliver, M., & Pous, R. (2018). CUIDATS: An RFID–WSN hybrid monitoring system for smart health care environments. *Future Generation Computer Systems*, *78*, 602–615. https://doi.org/10.1016/j.future.2016.12.023

21) Singh, N., Sasirekha, S. P., Dhakne, A., Thrinath, B. V. S., Ramya, D., & Thiagarajan, R. (2022). IOT enabled hybrid model with learning ability for E-health care systems. *Measurement: Sensors*, *24*(November), 100567. https://doi.org/10.1016/j.measen.2022.100567