Received January 20, 2022, accepted January 27, 2022, date of publication February 1, 2022, date of current version February 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3148600

Machine Learning Algorithms for COPD Patients Readmission Prediction: A Data Analytics Approach

ISRAA MOHAMED[1,2], MOSTAFA M. FOUDA[3], (Senior Member, IEEE), AND KHALID M. HOSNY[2], (Senior Member, IEEE)

[1]Faculty of Engineering and Computer Sciences, King Salman International University, Ras Sedr, Egypt
[2]Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt
[3]Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID 83209, USA

Corresponding author: Khalid M. Hosny (k_hosny@yahoo.com)

ABSTRACT Patients' readmission can be considered as a critical factor affecting cost reduction while maintaining a high-quality treatment of patients. Therefore, predicting and controlling patients' readmission rates would significantly improve the healthcare service. In this study, we aim at predicting the readmission of COPD (Chronic Obstructive Pulmonary Disease) patients through the deployment of machine learning algorithms. Area Under Curve (AUC) and ACCuracy (ACC) were considered as the main criteria for evaluating models' prediction power in each time frame. Then, the importance of the variables for each outcome was explicitly identified, and defined important variables have then been differentiated. Our study could achieve the highest accuracy in predicting readmission with %91 ACC.

INDEX TERMS Classification algorithms, COPD readmission, data mining, decision support systems, healthcare data analytics.

I. INTRODUCTION

Predictive analytics is one of the most commonly used IT (Information Technology) techniques in healthcare. For example, the possibility of using electronic health records as a basis for healthcare analysis for smart health was discussed by [1], [2]. Predictive analytics and data mining have also been utilized by [3] to control the propagation of chronic diseases. Generally, during the recent decade, healthcare-related research has focused on developing and implementing IT models to address the specific and critical needs of healthcare systems [4]. Most of these studies focus on utilizing big amounts of data to obtain valuable information and insights about the current and future behavior of the system under consideration.

Chronic Obstructive Pulmonary Disease (COPD) can be defined as a lung disease recognized by airflow fettering [5]. Worldwide, COPD has been considered as one of the major causes leading to higher rates of death. The Global Burden of Disease Study estimated 251 million spread cases of COPD in 2016. It was also reported that 3.17 million deaths were caused by COPD in 2015 (i.e., 5% of all deaths in that year), [6]. The admission rate of COPD patients rate in the United Kingdom has been doubled between 1991 and 2000 and by 2000, reported 1% of all hospital admissions [7]. The total costs of lung diseases in the EU (European Union) have been estimated to be about 6% of the total healthcare costs, and COPD was reported as taking the most significant percentage (%56) of these costs [5].

Readmission can be considered as one of the significant issues facing any healthcare system and one of the main causes of declined health services. Readmission can be defined as admitting patients to the hospital within a maximum of 30 days after being discharged from the same hospital earlier. Hospital readmission is costly for patients and hospitals as well [8]. Consequently, hospitals are striving to make sure that patients will receive adequate treatment in their first admission to minimize the possibility of readmission.

In this study, we are utilizing the most powerful data mining techniques such as Decision Trees (DT), Artificial Neural Networks (ANN), and Support Vector Machine (SVM) to

The associate editor coordinating the review of this manuscript and approving it for publication was Szidónia Lefkovits.

develop classification models that can determine the targeted group of high-risk COPD patients who are most likely to be readmitted to the hospital within 30 days of their discharge.

Unplanned readmission has been approached by many researchers. However, there is still an apparent lack of proof of their effectiveness [9]. One of the suggested causes of those studies' inefficiency may be attributed to their wasted work hunting the wrong targeted group of patients (i.e., patients with low risk of readmission) [10]. Therefore, there is a high need for reliable predictive models that are capable of accurately identifying high-risk patients most efficiently, allowing healthcare stakeholders to respond accordingly.

To the best of our knowledge, the multitude of algorithms such as SVM with multiple different Kernel functions has not been applied in the same context before. This study addresses a classification problem with two main target classes, namely readmitted and non-readmitted patients over a specific time frame. The rest of this paper is organized as follows: Section 2 reviews the related literature. The methodology applied and the proposed models are presented in Section 3. Section 4 displays the results. Finally, section 5 concludes the paper with further discussions of the results.

II. RELATED WORK

As we mentioned before, readmission may be defined as admitting patients to the hospital after a short time from their discharge. This short time has been set in the literature to be within 30 to 90 days [11], [12]. In this study, we set our readmission time frame to be within 30 days, as normally, healthcare service quality is measured by death rates within 30 days of discharge [13], [14]. Hospitals' readmission research is usually based on variables and data sets for a particular population, patient type, or specific disease because of the complex data collection procedures required to get a large amount of data. However, enough amount of data has a significant implication on the precision and accuracy of the developed predictive models. In this study, we are enhancing the generality of the developed predictive model through a large amount of data (around 620,000 entries). It sometimes happens that different classes are not equally represented which is referred to in the literature as the class imbalance problem [15]. Since most of the diseases are not usually found in the whole population, the class imbalance problem may be considered a common problem in the healthcare services field [16]. Undoubtedly, predictive analytic models are highly affected by the class imbalance problem [17]. Therefore, the developed classification models must take this problem into account and apply some compensation techniques. The most commonly used compensation techniques to balance classes are the different error cost negatives technique [18], the over-sampling technique [19], and the under-sampling technique [20]. To the best of our knowledge, there are limited studies in the literature that present the problem of imbalanced readmission data [21]. Another critical problem that arises when attempting to predict hospital readmission is the cost imbalance misclassification problem [22]. The cost imbalance problem is usually related to the above-mentioned class imbalance problem, and hence solution techniques of both problems can enhance each other [15]. To the best of our knowledge, the readmission predictive literature has rarely considered the cost imbalance misclassification problem. Machine learning algorithms have been widely used in the literature to classify readmitted patients. Most commonly used algorithms are: Logistic Regression (LR), Naïve Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest (RF) [23], [24], and [25]. Different studies compared predictive models based on their predicted output [23], [25], and [26]. However, most of these studies suffer from poor prediction quality, as the AUC ranged from 0.57 to 0.74, with only one excepted study of [27], who reported an AUC value of 0.83. So, the low prediction capability may be added to the challenges of developing predictive hospital readmission models.

Although COPD is considered a sserious disease with complicated consequences, it received little attention from researchers. The available literature studying the risk factors affecting COPD patients' admission and readmission is rare. On the other hand, many studies are focusing on these factors for other classes of patients. For example, [28] predicted the risk of heart failure patients' readmission using a multi-layer approach. They examined if patients will ever be readmitted or if they will be readmitted within short (30 days) or long (60 days) readmission time. Naïve Bayes and Support Vector Machine was the applied classification algorithms in their study. In 2017, [29] predicted the risk of heart failure patients' readmission using the NB classification algorithm. The highest reached accuracy of their model based on ACC and AUC was around 85% and 0.77, respectively. Reference [30] predicted the risk factors of heart failure patients' readmission based on a 30-day time horizon. They applied the LR classification model and reached an accuracy of 0.78 based on the AUC measure. The work done in [31], [32], and [33] are of the rare studies that approached COPD patients. Binson *et al.* [31] attempted to early diagnose COPD, lung cancer, and asthma through the utilization of an electronic nose, which analyzes human exhaled breath and classifies it according to different machine learning models. Their results achieved high levels of accuracy for the three diseases. Dhar proposed a novel ensemble model for the early detection of COPD [32]. The authors adopted 8 classifiers arranged in 2 different pools. A genetic algorithm has been utilized to find optimal hyper-parameters for each classifier. The results of their model outperform most of the recent Machine Learning models applied for COPD early detection. Wu *et al.* [33] considered the problem of readmission prediction for COPD patients using a novel CORE (COPD – Readmission) score, which predicts patient's readmission based on five main predictors, i.e., eosinophil count, lung function, triple inhaler therapy, previous hospitalization, and neuromuscular disease. It was found that there is a high correlation between the

CORE score and the COPD readmission, where a high CORE score meant a high risk of readmission and a short time to readmission.

Unplanned readmission may be attributed to different reasons, such as premature discharge, limited social service support [34], complications associated with the previous disease, and retrogression of initial health condition [9]. Other factors related to patients themselves maybe also of great importance, such as bad self-care and medication problems [34], [9]. Healthcare services may be measured by the level of unplanned readmission [21]. Higher rates of unplanned readmission indicate limited clinical management, which will reveal its consequences in hospitals in the long run. In this study, we are aiming at understanding risk factors related to admission and readmission of COPD patients in an Egyptian private hospital. Data has been collected from Al-Ghandoor Hospital (GH), Ash-Sharkia, Egypt. GH is considered the biggest private hospital in the city, providing more than 85% of total health services for the city population. The emergency department is the first stop for COPD patients with early symptoms. High-risk patients are then admitted to the hospital to receive appropriate health services. In this study, all COPD admissions to GH from January 2019 to December 2019 are included. Calculation of unplanned readmission rates of COPD patients and identifying risk factors leading to this unplanned readmission are the two main objectives of our study.

We approached the problem from different aspects than was previously done. Firstly, a conductive data collection phase was performed to gather accurate data that led to valid results. Then, a multitude classification algorithm (SVM with different Kernel functions) was applied. Two-time frames were used for the target class rather than a one-time frame. Our study could achieve the highest accuracy in predicting readmission with %91 ACC. This study was considered as a foundation step toward building a highly accurate predictive system.

III. METHODOLOGY

Data mining techniques have been applied in [35] to predict survival rates of heart transplant patients. The authors have used a data analytic methodology based on four phases. In this study, we are using the same methodology with slight changes to be more customized with our study scope and objectives. The general framework is presented in Figure 1. The data preparation phase (1st phase) is composed of three main steps: 1) specification of source of data, 2) cleaning of data (removing factors with poor prediction capability, clearing missing data, and eliminating faulty records), 3) selection of data (assigning a binary number to each readmitted patient indicating whether h/she has been readmitted within 1 or 3 months). Classification models are then applied to clean data in the second phase. The output of the second phase is a list from each classification model of important variables ranked based on their prediction power. Classification models accuracy is then assessed in the third phase. In the last phase of the methodology, important variables are compared based on their change over time.

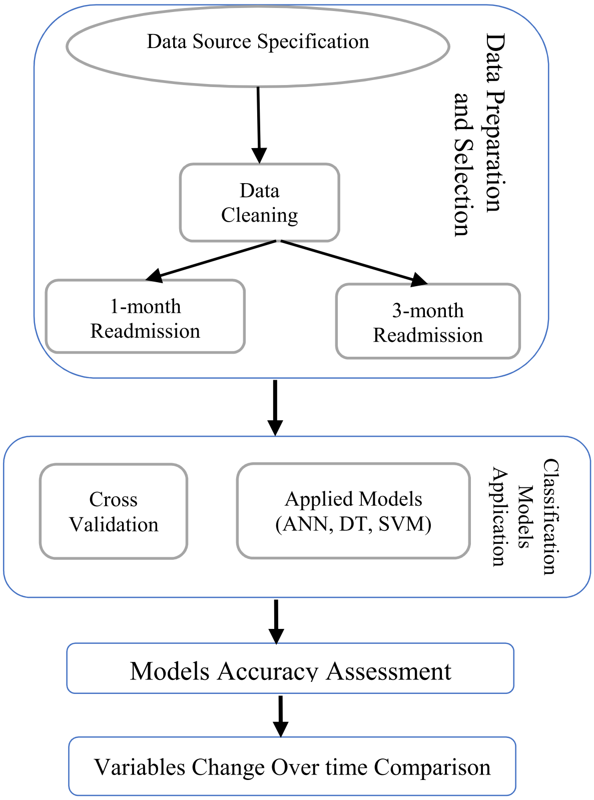
Data Source Specification
Data Preparation and Selection
Data Cleaning
1-month Readmission
3-month Readmission
Cross Validation
Applied Models (ANN, DT, SVM)
Classification Models Application
Models Accuracy Assessment
Variables Change Over time Comparison

FIGURE 1. Methodology general framework.

A. DATA PREPARATION

Data for this study has been collected from GH for the calendar year 2019. Data records include admittance and discharge information for all COPD inpatients in GH. Data has been collected from the moment of patients' admission and throughout their hospital stay and until 3 months from their

discharge. After the patient's admission, a complete record of data is generated, including admission data, medical history, laboratory results (CBC, BUN), and primary and secondary diagnosis. Most of the data were collected from the hospital information system, while the rest was based on direct observations and short interviews. Collected data is then presented to a pulmonologist and a Respiratory therapist to verify it. To achieve one of the study objectives, which is predicting patients' readmission, we needed to find the reasons of patients were readmitted from collected data. The collected data contain around 198 records of COPD patients' data.

Data cleaning includes data filtration, data excluding, and data merging. Filtration has been applied to poor data (i.e. data with low quality or limited power). Merging has been applied to records of the same patient if h/she suffered from having more than one admission on a single day to the same medical specialty. Excluding has been applied to

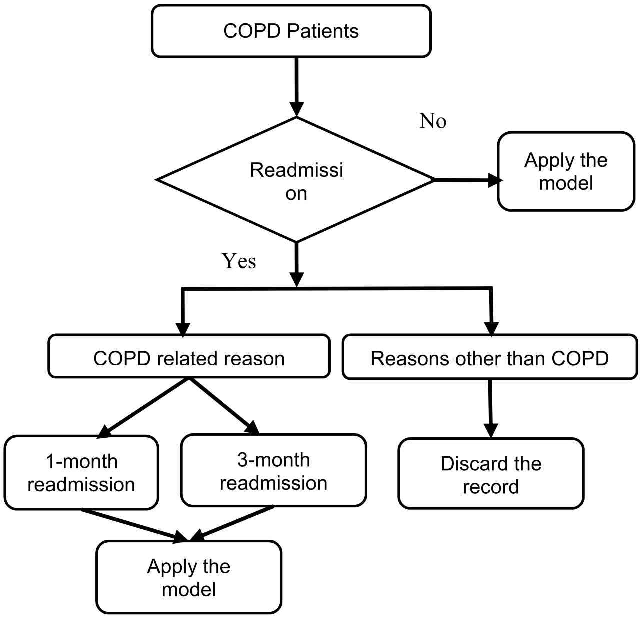

FIGURE 2. Data selection.

data not directly related to our scope of studies, such as age inconsistencies, patient registry number, erroneous data, records with "NA" term, and inconsistent components. After the collection and preparation of the data set, we ended up with 195 records and 32 variables, which were then divided into 3 parts according to the ratio 2:2:1, respectively. The first part was the training part of data, the second part was the validation part, and the last part was for testing.

Finally comes the selection of data step. In this step, only readmitted COPD patients were selected. This step is further illustrated in Figure 2.

B. CLASSIFICATION MODELS

Three main machine learning algorithms have been applied in the current study, namely: Artificial Neural Networks (ANNs), Decision Trees (DTs), and Support Vector Machines (SVMs). These algorithms have been chosen based on previous studies' recommendations on their satisfying performance [36]–[39], and [40]. The following subsections provide a brief description of each of the mentioned algorithms.

1) ARTIFICIAL NEURAL NETWORKS (ANNS)

An artificial neural network (ANN) is the part of computing systems prepared to imitate the human brain. It is the basis of AI and able to solve problems that normal humans find difficult to solve. ANNs have self-learning capabilities that enable them to produce better results as more data becomes available. The ANNs mimic the human brain in its general structure which consists of connected nodes responsible for processing and transmitting information to and from the brain. ANNs have been applied in many areas such as E-mail services, optimization problems, E-commerce, clustering and categorization, pattern recognition, prediction and forecasting, and deep learning techniques. ANNs can learn automatically from examples which makes them more attractive than other conventional artificial intelligence techniques.

2) DECISION TREES (DTS)

Decision Trees (DTs) are considered the most potent and popular classification technique. DTs are flowcharts taking the structure of an upside-down tree. Tree nodes represent tests on some specific characteristic, tree branches represent test results, and tree leaves represent class labels. DT classifier construction does not need domain knowledge which makes it suitable for discovering preparatory knowledge. DTs have many remarkable characteristics that make them superior among different classifying techniques. It can handle large amounts of complex structured data. Its classifier has a high level of accuracy, and its induction is capable of learning knowledge during the classification process. DT also has its own generated rules for prediction, making the model interpretation easier, more obvious, and coherent. DTs have been widely used recently as a data mining and machine learning technique to predict various system behaviors which makes it very suitable to predict COPD patients' readmission. Different variations of DTs algorithms can be found in the literature. For example, [41] uses C4.5, C5, and ID3 in his study. DTs algorithms that have been used in our research are namely: CHAID, C5, and C&RT.

3) SUPPORT VECTOR MACHINES (SVMS)

Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for regression analysis, classification, and predictive model development. In a classification problem, we generally have n number of features needed to be classified. In SVM, an n-dimensional space is plotted, and each feature value is represented by a specific coordinate value, while each data item is represented by a particular point in the feature space. Classification is then performed by determining the hyper-plane that best distinguishes the two classes. The SVM classifier is the border that separates the two classes in the best possible way. SVM classifier can be applied to both linear and non-linear datasets. Non-linear data sets can be transformed into linear by applying some kernel functions [42]. In this study, the applied kernel functions are namely: the sigmoid, the radial basis, and the polynomial functions.

IV. MODEL ACCURACY ASSESSMENT

Generally, in machine learning applications, models can't be fitted on the training data, and it cannot be said with a high degree of accuracy whether the model will work for the real data or not. For this reason, we need to ensure that our model got the correct patterns from the data. For this purpose, the cross-validation technique was employed.

TABLE 1. Confusion matrix.

True Label	Predicted Label	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Cross-validation is a technique in which models are trained using subsets of the dataset and then evaluated using the complementary subsets of the dataset. Three main steps are involved in cross-validation, which are: 1) Reserving a subset of the data, 2) Using the rest of the data set to train the model, 3) Testing the model using the reserved subset of data. Cross-validation techniques have many methods; the most commonly used method is the k-fold cross-validation. In the k-fold cross-validation, the data set is split into k subsets (folds). Training is then performed on all subsets except one (reserved subset), which is then used in testing the model. The method is iterated k times with different reserved subsets for each iteration. In this study, we applied a 5-cross validation approach using IBM SPSS Modeler version 17.

V. EXPERIMENTAL RESULTS

AUC and ACC for all the mentioned classification models have been computed for the two-time frames (1-month and 3-months) as represented in Table 2. These measures can be estimated based on the confusion matrix entries (see Table 1).

Accuracy (ACC) is calculated using the formula

$$\mathrm{ACC} = \frac{TN + TP}{TN + TP + FN + FP}$$

The Area Under the Curve (AUC) is a measure of the ability of a classifier to differentiate between classes and is used as a summary of the Receiver Operator Characteristic (ROC) curve. ROC is a probability curve that plots the True Positive (TP) rate against the False Positive (FP) rate at various threshold values. The higher the AUC, the better the performance of the model at differentiating between the positive and negative classes. As illustrated in Figure 3 (a), When AUC = 1, then the classifying model can perfectly distinguish between all the Positive and the Negative points correctly. If, on the other hand, the AUC had been 0, then the classifying model would be predicting all Negatives as Positives and all Positives as Negatives. However, as represented in Figure 3 (b), when AUC falls between 0.5 and 1, then there is a high probability that the classifying model will differentiate the positive class values from the negative class values. This is so because the classifier can detect more numbers of true positives and true negatives than false negatives and false positives.

Precision is calculated using the formula:

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

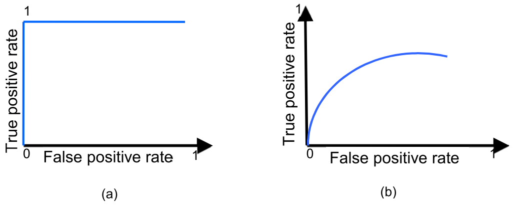

FIGURE 3. AUC interpretation.

F1 score is calculated using the formula:

$$F1 = \frac{2TP}{(TP + FP) + (TP + FN)}$$

IBM SPSS Modeler was the software used for calculating the measurement criteria. AUC has proved to be more robust than ACC, and so it has been used as the basis for sorting Table 2. Each performance measure for classification models is associated with its precision and F1-score, as presented in Table 2.

As illustrated in Table 2, the performance measures values are higher in the 1-month time frame than those in the 3-month time frame, making them more credible. The maximum AUC value of the 1-month time frame models is 0.77 (CHAID and exhaustive CHAID Tree Algorithms), while the maximum AUC value of the 3-month time frame models is 0.64 (ANN-RBF model). It was also noticed that the 1-month time frame models experienced higher ACC values than its counterpart in the 3-month time frame. For example, the ACC value of the C5 model in the 1-month time frame is 89.9% (maximum ACC value in this timeframe), while its value in the CHAID and exhaustive CHAID models in the 3-month time frame is 67.7% (maximum ACC value in this time frame). The high values of the performance measures in the 1-month time frame can be justified by two possible reasons. It might be attributed to the fact that patients receive their medication and further treatment through continuous follow-ups, which gives patients in the longer time frame more time to finish their treatment after being discharged. It might also happen because of the limited archived patients' data, which indicates that the models need more patients' records to be able to achieve better readmission prediction performance in the case of the 3-month time frame.

Factor importance has been calculated for each variable in each model for both of the time frames. The variable whose factor importance is higher than 0.00 is considered as an important variable for that model. Each model has its own set of important variables, while a variable may be important in one model and unimportant in another model. For example, the gender variable is considered an important variable in the C5 model, while the age variable is considered to be important in the ANN-MLP and RBF models. We have divided the 32 variables into six groups and monitored their importance change over time according to the number of models they are considered important

TABLE 2. Performance measures for each model in the two-time frames.

Model	1-Month Time Frame				Model	3 – Month Time Frame			
	ACC	AUC	Precision	F1 Score		ACC	AUC	Precision	F1 Score
Tree-As (CHAID)	88.7	0.77	0	0/0	ANN – RBF	64.8	0.64	0.15	0/0
Tree-As (Exhaustive CHAID)	88.7	0.77	0	0/0	Tree-As (CHAID)	67.7	0.62	0.20	0/0
CHAID	89.1	0.75	0.31	0.41	Tree-As (Exhaustive CHAID)	67.7	0.62	0.17	0/0
SVM– Polynomial	88.3	0.73	0.11	0/0	ANN – MLP	66.1	0.61	0.14	0.20
SVM – RBF	87.4	0.73	0	0/0	C5	59.1	0.48	0.25	0.26
C5	89.9	0.72	0.42	0.48	SVM – RBF	60.0	0.47	0.11	0.14
ANN – MLP	64.8	0.68	0.11	0/0	CHAID	53.0	0.46	0.24	0.25
ANN – RBF	87.4	0.67	0.16	0/0	SVM– Sigmoid	66.8	0.46	0	0/0
SVM – Linear	81.3	0.63	0.15	0.17	SVM– Polynomial	53.0	0.45	0.23	0.24
SVM-Sigmoid	88.7	0.38	0	0/0	SVM– Linear	54.3	0.39	0.08	0/0

TABLE 3. Variables distribution and assessing.

Group	NMI	Average Importance (1-month timeframe)	Average Importance (3-month time frame)
1	10	0.316	0.012
2	8, 9	0.178	0.294
3	6, 7	0.021	0.003
4	5	0.198	0.247
5	3, 4	0.005	0.008
6	0, 1, 2	0.0004	0.00005

in (NMI). Table 3 illustrates the distribution and assessment of the variable.

Variables change over time is more illustrated in Figure 4. Variables were ordered in groups according to their importance in predicting readmission. However, the importance of the variables is not the same for the two-time frames. For example, the most important variable in predicting readmission in the 1-month time frame is the Low Lung Function (LLF), which is included in group 1 with mean importance of 0.316 across the 10 applied models. The LLF importance for the 3-month time frame is much lower than its importance in the 1-month time frame with mean importance of 0.012 across the 10 models. The Beck Anxiety Inventory (BAI) scores variable is included in group 2 with mean importance of 0.178 in the 1-month time frame and higher mean importance of 0.294 in the 3-month time frame. The BAI has been considered important in 8 models. Group 3 includes Uric Acid (UA) and Base Creatinine (BC) variables. While the UC variable was considered important in the 7 models, the BC was considered important in 6 models. The UC mean importance is higher in the 1-month time frame (0.021) than its counterpart in the 3-month time frame (0.003). On the other hand, the BC has almost the same importance in both time frames with mean importance of 0.031 in the 1-month time frame and 0.029 in the 3-month time frame. In group 4, Cardiovascular Diseases (CD) are considered the most important factor in predicting the 3-month time frame readmission, with mean importance of 0.247 and 0.198 the 1-month time frame. Group 5 contains variables that were considered important in the 4 model or less. For example, sex and age variables have been considered important in 4 and 3 models, respectively. Variables that were considered important in the two models at most were grouped in group 6. Comorbidities, Beck Depression Inventory (BDI), and decreased physical activity have different importance factors for each time frame. For example, the decreased physical activity shows a large loss of importance from the 3-month time frame to the 1-month time frame (around 48% loss of importance).

VI. DISCUSSION

This paper developed a data-driven model to predict readmission of COPD patients to hospitals within one 1-month or 3-months from their discharge. The paper also studied variables' prediction importance and their change over time. A data mining approach has been utilized to achieve the study aims. The general framework of our methodology is composed of four phases: data preparation phase (1st phase), data cleaning phase (2nd phase), classification models accuracy assessment phase (3rd phase), and variables comparison phase (4th phase).

Our methodology has been applied to data collected from GH for the calendar year 2019. Data records include admittance and discharge information for all COPD inpatients in GH. Data has been collected from the moment of patients' admission and throughout their hospital stay and until 3-months from their discharge. Collected data contained around 198 records of COPD patients' data. Excluding has been applied to data not directly related to our scope of

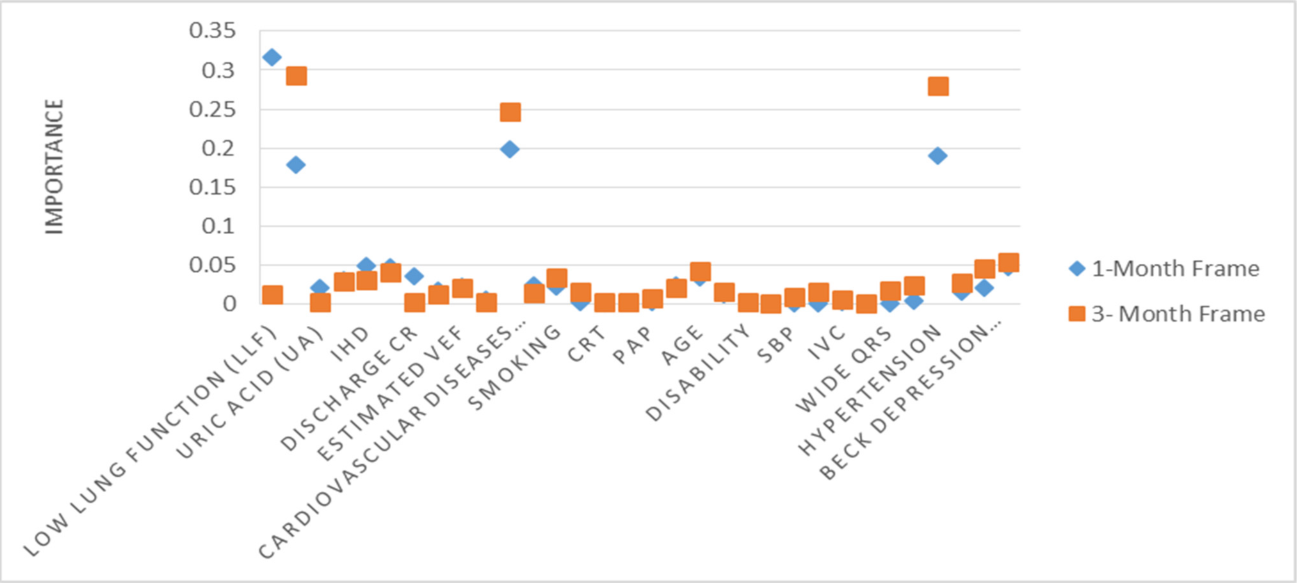

FIGURE 4. Variables change over time.

studies, such as age inconsistencies, patient registry number, erroneous data, records with "NA" term, and inconsistent components. After the collection and preparation of the data set, we ended up with 195 records and 32 variables, which were then divided into three parts according to the ratio 2:2:1, respectively. Our study aimed at answering some research questions such as: will it be possible to distinguish higher importance factors that contribute more to the prediction of COPD patients' readmission? How do these factors' importance change over time (1-month and 3-month time frames)? Which of the applied classification algorithms is more powerful in predicting COPD patients' readmission?

Our methodology could predict readmission within 1-month and 3-month timeframes with an average AUC score of 0.68 and 0.52, respectively. The average ACC score was 85.43 and 61.25, respectively. Hence, the methodology is more reliable in predicting readmission of COPD patients within one 1-month from their discharge. However, for longer time frames (3-month or more), it is not so reliable and needs more data records to build more reliable and powerful models. Furthermore, our methodology could define different important factors of the variables and identify their change over time.

VII. CONCLUSION

Our main contribution can be summarized as the use of machine learning algorithms and techniques to handle the class imbalance problem utilizing medical vector scattering to cope with the limited conventional readmission predictive models and hence improving predictability. We compare the different machine learning algorithms according to their predictability power of hospital readmission prediction.

Nevertheless, our study still has some limitations due to the self-funding and limited budget. It was also very challenging to have a team of experts dedicated to our study needs (data collection, cleaning, and preparation). Another important point that limited our study was the limited number of COPD patients' records (195 records).

Our future research direction is to study and investigate alternative classification techniques to further amend the classification model. We also plan to study more detailed predictions on hospital readmission, which has a higher effect in designing and building more efficient and effective post-discharge models. For example, the probability of patient readmission within a specific time point is interesting to predict and study how this probability may be affected by earlier hospitalization events [43], [44].

ACKNOWLEDGMENT

The authors would like to thank the medical and management team in GH for their support in data collection and preparation. The authors are also grateful to the anonymous reviewers and editorial board for their many insightful comments, which have significantly improved this article.

REFERENCES

[1] K. Yang, X. Li, H. Liu, J. Mei, G. Xie, J. Zhao, B. Xie, and F. Wang, "TaGiTeD: Predictive task guided tensor decomposition for representation learning from electronic health records," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2017, pp. 2824–2830.

[2] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson, and H. Liu, "Deep learning and alternative learning strategies for retrospective real-world clinical data," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–5, Dec. 2019, doi: 10.1038/s41746-019-0122-0.

[3] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, 2016, pp. 381–386, doi: 10.1109/ICATCCT.2016.7912028.

[4] I. Mohamed, "A discrete event simulation model for waiting time management in an emergency department: A case study in an Egyptian hospital," *Int. J. Model., Simul., Sci. Comput.*, vol. 12, no. 1, Feb. 2021, Art. no. 2050063, doi: 10.1142/S1793962320500634.

[5] D. Singh *et al.*, "Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: The GOLD science committee report 2019," *Eur. Respiratory J.*, vol. 53, no. 5, May 2019, Art. no. 1900164, doi: 10.1183/13993003.00164-2019.

[6] R. Rodriguez-Roisin, K. F. Rabe, J. Vestbo, C. Vogelmeier, and A. Agustí, "All previous and current members of the science committee and the board of directors of GOLD (goldcopd.Org/committees/). Global initiative for chronic obstructive lung disease (GOLD) 20th anniversary: A brief history of time," *Eur. Respiratory J.*, vol. 50, no. 1, Jul. 2017, Art. no. 1700671, doi: 10.1183/13993003.00671-2017.

[7] J. Menzin, L. Boulanger, J. Marton, L. Guadagno, H. Dastani, R. Dirani, A. Phillips, and H. Shah, "The economic burden of chronic obstructive pulmonary disease (COPD) in a U.S. medicare population," *Respiratory Med.*, vol. 102, no. 9, pp. 1248–1256, Sep. 2008, doi: 10.1016/j.rmed.2008.04.009.

[8] M. Chen and D. C. Grabowski, "Hospital readmissions reduction program: Intended and unintended effects," *Med. Care Res. Rev.*, vol. 76, no. 5, pp. 643–660, Oct. 2019, doi: 10.1177/1077558717744611.

[9] L. O. Hansen, R. S. Young, K. Hinami, A. Leung, and M. V. Williams, "Interventions to reduce 30-day rehospitalization: A systematic review," *Ann. Internal Med.*, vol. 155, no. 8, pp. 520–528, 2011, doi: 10.7326/0003-4819-155-8-201110180-00008.

[10] A. Steventon, M. Bardsley, J. Billings, T. Georghiou, and G. Lewis, *An Evaluation of the Impact of Community-Based Interventions on Hospital Use*. London, U.K.: Nuffield Trust, 2011.

[11] I. Eigner and A. Cooney, "A literature review on predicting unplanned patient readmissions," in *Delivering Superior Health and Wellness Management With IoT and Analytics. Healthcare Delivery in the Information Age*, N. Wickramasinghe and F. Bodendorf, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-17347-0_12.

[12] C. M. Ashton and N. Wray, "A conceptual framework for the study of early readmission as an indicator of quality of care," *Social Sci. Med.*, vol. 43, 11, pp. 1533–1541, 1996, doi: 10.1016/s0277-9536(96)00049-4.

[13] S. Mulpuru, J. McKay, P. Ronksley, K. Thavorn, D. Kobewka, and A. J. Forster, "Factors contributing to high-cost hospital care for patients with COPD," *Int. J. Chronic Obstructive Pulmonary Disease*, vol. 12, pp. 989–995, Mar. 2017, doi: 10.2147/COPD.S126607.

[14] T. Wen, B. Liu, X. Wan, X. Zhang, J. Zhang, X. Zhou, A. Y. L. Lau, and Y. Zhang, "Risk factors associated with 31-day unplanned readmission in discharged patients after stroke in China," *BMC Neurol.*, vol. 18, no. 1, pp. 1–11, 2018, doi: 10.1186/s12883-018-1209.

[15] M. Gary Weiss, K. Mccarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs," *Dmin*, vol. 7, p. 24, Jun. 2007.

[16] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, pp. 2–3, pp. 427–436, 2008, doi: 10.1016/j.neunet.2007.12.031.

[17] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, "Predictive modeling of hospital readmissions using metaheuristics and data mining," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7110–7120, Nov. 2015, doi: 10.1016/j.eswa.2015.04.066.

[18] E. Vittinghoff, V. D. Glidden, C. S. Shiboski, and E. C. McCulloch, *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Cham, Switzerland: Springer, 2011.

[19] V. N. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*, 2009, pp. 875–886, doi: 10.1007/0-387-25465-X_40.

[20] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2008, doi: 10.1109/TSMCB.2008.2007853.

[21] D. Golmohammadi and N. Radnia, "Prediction modeling and pattern recognition for patient readmission," *Int. J. Prod. Econ.*, vol. 171, pp. 151–161, Jan. 2016, doi: 10.1016/j.ijpe.2015.09.027.

[22] H. K. Lee, R. Jin, Y. Feng, P. A. Bain, J. Goffinet, C. Baker, and J. Li, "An analytical framework for TJR readmission prediction and cost-effective intervention," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1760–1772, Jul. 2019, doi: 10.1109/JBHI.2018.2859581.

[23] E. Demir, "A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines," *Decis. Sci.*, vol. 45, no. 5, pp. 849–880, Oct. 2014, doi: 10.1111/deci.12094.

[24] A. Agarwal, C. Baechle, R. Behara, and X. Zhu, "A natural language processing framework for assessing hospital readmissions for patients with COPD," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 2, pp. 588–596, Mar. 2018, doi: 10.1109/JBHI.2017.2684121.

[25] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," *J. Biomed. Informat.*, vol. 56, pp. 229–238, Aug. 2015, doi: 10.1016/j.jbi.2015.05.016.

[26] P. C. Austin, "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality," *Statist. Med.*, vol. 26, no. 15, pp. 2937–2957, Jul. 2007, doi: 10.1002/sim.2770.

[27] E. A. Coleman, S.-J. Min, A. Chomiak, and A. M. Kramer, "Posthospital care transitions: Patterns, complications, and risk identification," *Health Services Res.*, vol. 39, no. 5, pp. 1449–1466, Oct. 2004, doi: 10.1111/j.1475-6773.2004.00298.x.

[28] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, and B. Krishnapuram, "Predicting readmission risk with institution-specific prediction models," *Artif. Intell. Med.*, vol. 65, no. 2, pp. 89–96, Oct. 2015, doi: 10.1016/j.artmed.2015.08.005.

[29] K. Shameer, K. W. Johnson, A. Yahi, R. Miotto, L. Li, D. Ricks, J. Jebakaran, P. Kovatch, P. P. Sengupta, S. Gelijns, A. Moskovitz, B. Darrow, D. L. David, A. Kasarskis, N. P. Tatonetti, S. Pinney, and J. T. Dudley, "Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using Mount Sinai heart failure cohort," in *Proc. Biocomputing*, Jan. 2017, pp. 276–287, doi: 10.1142/9789813207813_0027.

[30] K. T. G. Leong, L. Y. Wong, K. C. Y. Aung, M. Macdonald, Y. Cao, S. Lee, W. L. Chow, S. Doddamani, and A. M. Richards, "Risk stratification model for 30-day heart failure readmission in a multiethnic south east Asian community," *Amer. J. Cardiol.*, vol. 119, no. 9, pp. 1428–1432, May 2017, doi: 10.1016/j.amjcard.2017.01.026.

[31] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew, "Prediction of pulmonary diseases with electronic nose using SVM and XGBoost," *IEEE Sensors J.*, vol. 21, no. 18, pp. 20886–20895, Sep. 2021.

[32] J. Dhar, "Multistage ensemble learning model with weighted voting and genetic algorithm optimization strategy for detecting chronic obstructive pulmonary disease," *IEEE Access*, vol. 9, pp. 48640–48657, 2021.

[33] Y.-K. Wu, C.-C. Lan, I.-S. Tzeng, and C.-W. Wu, "The COPD-readmission (CORE) score: A novel prediction model for one-year chronic obstructive pulmonary disease readmissions," *J. Formosan Med. Assoc.*, vol. 120, no. 3, pp. 1005–1013, Mar. 2021.

[34] S. Chen, N. Kong, X. Sun, H. Meng, and M. Li, "Claims data-driven modeling of hospital time-to-readmission risk with latent heterogeneity," *Health care Manage. Sci.*, vol. 22, no. 1, pp. 156–179, 2019, doi: 10.1007/s10729-018-9431-0.

[35] A. Dag, A. Oztekin, A. Yucel, S. Bulur, and F. M. Megahed, "Predicting heart transplantation outcomes through data analytics," *Decis. Support Syst.*, vol. 94, pp. 42–52, Feb. 2017, doi: 10.1016/j.dss.2016.10.005.

[36] S. G. Drakos, A. G. Kfoury, E. M. Gilbert, J. W. Long, J. C. Stringham, E. H. Hammond, K. W. Jones, D. A. Bull, M. E. Hagan, J. W. Folsom, B. D. Horne, and D. G. Renlund, "Multivariate predictors of heart transplantation outcomes in the era of chronic mechanical circulatory support," *Ann. Thoracic Surg.*, vol. 83, no. 1, pp. 62–67, Jan. 2007, doi: 10.1016/j.athoracsur.2006.07.050.

[37] K. N. Hong, A. Iribarne, B. Worku, H. Takayama, A. C. Gelijns, Y. Naka, V. Jeevanandam, and M. J. Russo, "Who is the high-risk recipient? Predicting mortality after heart transplant using pretransplant donor and recipient risk factors," *Ann. Thoracic Surg.*, vol. 92, no. 2, pp. 520–527, Aug. 2011, doi: 10.1016/j.athoracsur.2011.02.086.

[38] A. Oztekin, Z. J. Kong, and D. Delen, "Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations," *Decis. Support Syst.*, vol. 51, no. 1, pp. 155–166, Apr. 2011, doi: 10.1016/j.dss.2010.12.004.

[39] A. Kilic, E. S. Weiss, T. J. George, G. J. Arnaoutakis, D. D. Yuh, A. S. Shah, and J. V. Conte, "What predicts long-term survival after heart transplantation? An analysis of 9,400 ten-year survivors," *Ann. Thoracic Surg.*, vol. 93, no. 3, pp. 699–704, Mar. 2012, doi: 10.1016/j.athoracsur.2011.09.037.

[40] N. Nakayama, M. Oketani, Y. Kawamura, M. Inao, S. Nagoshi, K. Fujiwara, H. Tsubouchi, and S. Mochida, "Algorithm to determine the outcome of patients with acute liver failure: A data-mining analysis using decision trees," *J. Gastroenterol.*, vol. 47, no. 6, pp. 664–677, Jun. 2012, doi: 10.1007/s00535-012-0529-8.

[41] R. J. Quinlan, *C4. 5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1994.

[42] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *Morgan Kaufmann Ser. Data Manage. Syst.*, vol. 5, no. 4, pp. 83–124, 2011.

[43] J. E. Helm, A. Alaeddini, J. M. Stauffer, K. M. Bretthauer, and T. A. Skolarus, "Reducing hospital readmissions by integrating empirical prediction with resource optimization," *Prod. Oper. Manage.*, vol. 25, no. 2, pp. 233–257, Feb. 2016.

[44] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, and B. Krishnapuram, "Predicting readmission risk with institution-specific prediction models," *Artif. Intell. Med.*, vol. 65, no. 2, pp. 89–96, Oct. 2015.

ISRAA MOHAMED was born in Zagazig, Sharkia, Egypt, in 1985. She received the B.S. degree in information sciences and technology and the M.S. and Ph.D. degrees in operations research and decision support from the University of Zagazig, Egypt, in 2005, 2011, and 2017, respectively.

From 2005 to 2011, she was a Teaching Assistant with the Faculty of Computers and Informatics, Zagazig University. Since 2011, she has been an Assistant Lecturer with the Operations Research Department, Faculty of Computers and Informatics, Zagazig University. Since 2017, she has been a Lecturer with the Operations Research Department, Faculty of Computers and Informatics, Zagazig University. She is currently a Lecturer with the Faculty of Engineering and Computer Sciences, King Salman International University, South Sinai, Egypt. Her research interests include simulation and optimization applications in healthcare, machine learning, and data mining algorithms applications in healthcare systems.

MOSTAFA M. FOUDA (Senior Member, IEEE) received the Ph.D. degree in information sciences from Tohoku University, Japan, in 2011. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Idaho State University, ID, USA. He is also an Associate Professor with Benha University, Egypt. He has served as an Assistant Professor at Tohoku University. He was a Postdoctoral Research Associate with Tennessee Technological University, TN, USA. He has published more than 60 papers in prestigious peer-reviewed journals and conferences. His research interests include cybersecurity, communication networks, wireless mobile communications, smart healthcare, smart grids, AI, blockchain, and the IoT. He was a recipient of the prestigious first place award during his graduation from the Faculty of Engineering at Shoubra, Benha University, in 2002. He has served as the Symposium/Track Chair for the IEEE VTC2021-Fall Conference. He has also served as a the Workshops Chair, the Session Chair, a Technical Program Committee (TPC) Member, and a Designated Reviewer for leading international conferences, like IEEE GLOBECOM, ICC, PIMRC, ICCVE, IWCMC, and 5G World Forum. He also served as a Guest Editor for some special issues of several top-ranked journals, such as IEEE Wireless Communications and *IEEE Internet of Things Magazine*. He also serves as a referee of some renowned IEEE journals and magazines, such as IEEE Communications Surveys and Tutorials, IEEE Wireless Communications, IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Smart Grid, IEEE Access, IEEE Transactions on Network and Service Management, IEEE Transactions on Emerging Topics in Computing, and IEEE Network. He is an Editor of IEEE Transactions on Vehicular Technology and an Associate Editor of IEEE Access.

KHALID M. HOSNY (Senior Member, IEEE) was born in 1966, Zagazig, Egypt. He received the B.Sc., M.Sc., and Ph.D. degrees from Zagazig University, Egypt, in 1988, 1994, and 2000, respectively. From 1997 to 1999, he was a Visiting Scholar with the University of Michigan, Ann Arbor, and the University of Cincinnati, Cincinnati, USA. He is currently a Professor of information technology with the Faculty of Computers and Informatics, Zagazig University. He has published three edited books and more than 100 articles in international journals. His research interests include image processing, pattern recognition, multimedia, and computer vision. He is a Senior Member of ACM. He was listed among the top 1% of scientists according to the Stanford University rank, in 2020. He is an editor and a scientific reviewer for more than 50 international journals.