Recognition of Endotracheal Suctioning Activities: A Feature Extraction and Ensemble Learning Approach based on Pose Estimation Data This is Used Only for Thesis Report Purposes

1st Hoang Khang Phan
dept. of Biomedical Engineering
Ho Chi Minh city
University of Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
Khang.phan2411@hcmut.edu.vn

2nd Tu Nhat Khang Nguyen
dept. of Computer Science and Engineering
Ho Chi Minh city
University of Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
khang.nguyentusteven@hcmut.edu.vn

3rd Truong Vi Bui
dept. of Computer Science and Engineering
Ho Chi Minh city
University of Infomation Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
22210178@ms.uit.edu.vn

4th Khuong Cong Duy Nguyen
dept. of Computer Science and Engineering
Ho Chi Minh city
University of Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
duy.nguyenkhuongcong@gmail.com

5th Tuan Phong Nguyen
dept. of Biomedical Engineering
Ho Chi Minh city
University of Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
phong.nguyen0505@hcmut.edu.vn

6th Nhat Tan Le
dept. of Biomedical Engineering
Ho Chi Minh city
University of Technology-
Vietnam National University
Ho Chi Minh City, Vietnam
lenhattan@hcmut.edu.vn

Abstract—**Endotracheal suctioning is a critical yet intricate procedure in healthcare, often associated with potential complications due to its invasive nature. The escalating demand for certified professionals proficient in this technique, particularly in the context of home healthcare, underscores the necessity for comprehensive research. However, there exists a dearth of studies investigating automated recognition and analysis of nursing activities during endotracheal suctioning. This study utilizes skeleton pose data to establish correlations between nursing activities and pose dynamics, utilizing a dataset comprising 26 nurses and nurse students. Feature extraction encompassed a diverse array derived from pose characteristics, movement patterns, and activity sequences. For classification purposes, three ensemble learning pipelines were proposed, incorporating both original features and synthetic features generated by the Synthetic Data Vault model. The optimal pipeline, a voting classification approach, achieved noteworthy performance metrics, attaining 89.9% and 90.4% for overall weighted F1 score and accuracy, respectively. Furthermore, ANOVA f-test results revealed the significance of statistical features such as relative position and movement in activity discernment, with sequential features significantly augmenting model efficacy. Notably, disparities in activity execution were observed between experienced nurses and students, particularly evident in high-skill invasive maneuvers during endotracheal suctioning. These findings offer valuable insights for nurse training applications and highlight the potential of automated analysis in healthcare settings.**

If you want to see the full paper please go to this https://doi.org/10.1109/ABC61795.2024.10651640

979-8-3503-7550-3/24/$31.00 ©2024 IEEE

Index Terms—**Nurse care activity, Endotracheal Suctioning, Activity Recognition, Pose Estimation, Machine Learning.**

I. Introduction

Within the evolving realm of clinical healthcare, the practice of Endotracheal Suctioning (ES) emerges as a vital procedure for respiratory management, crucial for ensuring the patency of airways and mitigating the risk of respiratory distress. As the need for competent healthcare practitioners escalates, the integration of innovative educational strategies becomes increasingly significant. Among these innovations, the adoption of ES simulation systems for training stands out as a key advancement in healthcare education. These simulation platforms provide a controlled and safe setting, enabling learners to refine their clinical competencies through direct engagement. This methodology not only promotes uniform skill enhancement but also aids in the professional development of both established nurses and students pursuing nursing. Consequently, the development of a robust training simulation platform could serve as a crucial tool for healthcare workers lacking experience, offering them a valuable resource to enhance their skills effectively and efficiently.

To facilitate the analysis of task performance and to standardize Endotracheal Suctioning (ES) procedures, engaging in nurse care activities becomes essential. Within the framework of the ABC Challenge 2024, recognizing nurse care activities necessitates processing a dataset that includes video-recorded data of nurses performing nine distinct ES procedures, alongside skeleton data extracted from these recordings using YOLOv7 [2]. This research introduces a classification pipeline specifically designed for this dataset, with an emphasis on identifying variations among practitioners. The objectives of this study are specifically aimed at:

- Extracting a comprehensive set of features from pose-estimation data that capture the nuances of Endotracheal Suctioning (ES) activities. Such features hold promise for assessing the efficacy and proficiency of nurses' actions, thereby enabling robust evaluations of their efficiency and effectiveness.
- Proposing appropriate ensemble learning pipeline to optimize the recognition performance of Endotracheal Suctioning (ES) activities, thereby enabling accurate classification and a deeper understanding of procedural characteristics.
- Enhancing classification accuracy and addressing class imbalance by employing data augmentation techniques through generative modeling.
- Investigating the performance correlation between nurses with three years of experience and nursing students is essential to elucidate the influence of experience on procedural efficacy. Such findings hold the potential to deepen our comprehension of how experience levels affect the execution of procedures.

This research has the potential to significantly enhance healthcare training for nurses and improve hands-on skill performance by facilitating the use of simulation systems prior to clinical practice. Moreover, a deeper exploration of the ESTE-SIM [3] simulation system and the performances of nurses can provide crucial insights for future enhancements in Endotracheal Suction training for nursing students.

II. Related works

Skeleton pose analysis has become a pivotal field in computer vision and artificial intelligence, with broad applications in action recognition, human tracking, and human-machine interaction. Deep analysis of the skeleton pose can yield comprehensive information for recognizing activities. For example, Halim et al. extracted skeletal features by calculating all angles between any three joints and all distances between every pair of joints, organizing them into feature vectors for each static pose. Sequences of these feature vectors then represented complex dynamic actions. This methodology was applied to the most recent and largest benchmark, the MSRC-12 Kinect Gesture Dataset. By employing local and global random forests on the dataset, an impressive accuracy of 94.7% was achieved with one-third of the MSRC-12 dataset used for training, and an accuracy of 95.8% was achieved when training involved two-thirds of the dataset [4].

To overcome challenges associated with action diversity and skeletal variation, data augmentation methods alongside data preprocessing techniques, such as reshaping, have been extensively utilized to boost the performance of skeletal classification models. Despite the advancements in skeleton pose analysis, issues related to accuracy and processing speed remain, particularly in scenarios involving multiple subjects and occlusions. To improve the generalization capability of models for skeleton-based activity recognition, Meng et al. [5] designed a Sample Fusion Network (SFN) that combines Long Short-Term Memory (LSTM) and an autoencoder for data augmentation purposes. This approach enhanced the accuracy of the activity recognition model from 79.53% to 90.75% by applying the SFN model. In a different strategy to augment skeleton data, Hyilym Ramirez et al. [6] employed a Tabular Generative Adversarial Network (TABGAN) to synthesize feature frames extracted from skeleton data. Subsequently, classification was conducted on a dataset comprising both real and synthesized data using a Large Language Model, specifically the Bidirectional Encoder Representations from Transformers (BERT). The combination of BERT and TABGAN demonstrated a significant performance improvement, achieving 99.5% accuracy and an 87.2% F1-score, compared to 99.14% accuracy and an 80.95% F1-score with the BERT-only model. These studies highlight the effectiveness of data augmentation in addressing challenges related to interperson variance and dataset diversity. Therefore, selecting an appropriate data augmentation technique is crucial for the dataset used in this study.

In this study, we propose a comprehensive machine-learning pipeline designed to optimize the performance of ES activity recognition using skeleton-extracted data. The pipeline encompasses feature extraction strategies and a suitable data augmentation process employing a generative model. Additionally, this study investigates the performance of activities performed by nurses and students. The outcomes of this research will offer suggestions for future implementations and provide insightful analyses.

III. Methodology

A. Dataset

The dataset employed in this study originates from video recordings of Endotracheal Suctioning (ES) activities performed by ten experienced nurses, each with over three years of clinical suctioning expertise, and twelve nursing students from a university. These participants executed ES procedures using the ESTE-SIM simulation system. The dataset consists of two distinct types of data: video recordings captured from the front side of the participants, designated for training purposes, and Pose Skeleton (keypoints) data, extracted from the videos using YOLOv7, intended for both training and testing. The ES procedure is broken down into a total of nine distinct activities, with each activity assigned a unique identification number.

During the data collection phase, 22 participants executed the Endotracheal Suctioning (ES) procedure, which encompasses 9 activities (Catheter preparation, Temporal removement of an artificial airway, Suctioning phlegm, Refitting the artificial airway, Catheter disinfection, Discarding gloves, Positioning, Auscultation, and Others), on two occasions. Each recorded video features a frame rate of 30 frames per second

TABLE I
THE TABLE OF FEATURE TYPE AND QUANTITIES.

Feature group	Feature type	Quantities
Group A	Position and movement	2240
Group B	Relative position and movement	1920
Group C	Angle-based	180
Group D	Time-lag	4341
Group E	Sequential-based	2
	Total	8683

and an image resolution of 1920x1080 pixels. The dataset has been partitioned into a training set, comprising 32 videos, and a submission (or testing) set, consisting of 12 videos, with both sets including videos of nurses and students.

B. Data smoothing and segmentation

When the missing sequence extends up to 4 seconds, which event, from our analysis, was brought about by the technical failed to identify the body segment, a linear interpolation algorithm was employed to ensure continuity and accuracy in the skeletal data representation. Otherwise, if the missing sequence lasts longer than 4 seconds, the original values were retained. . After that, data was segmented by 3-second (or 90 frames) intervals for feature extraction.

C. Feature extraction

In this study, we employed 13 out of 17 body positions from the extracted skeleton data, encompassing: the nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left hip, right hip, left wrist, right wrist, left elbow, and right elbow, each tracked in both the x and y-axis dimensions.

From the YOLOv7 extracted skeleton key points, a total of 8681 features were extracted in this study, divided into 4 groups: Position and movement features; relative position and movement features; angle-based features; and time-lag features (as shown in Table I). Additionally, 2 sequential-based features were determined in this work based on the natural order of ES activity.

a) Position and movement features: Position and movement features encompass the coordinates of each body part and pose characteristics, including velocity, acceleration, and direction values. These features provide insight into how the subject executes the activity, such as the position of body parts, the movement speed, the pose force, or the direction of movement. By analyzing these aspects, it becomes possible to evaluate activity performance.

In this study, we examined the position and movement characteristics in the x and y coordinates of 14 marks (13 body marks from skeleton data and 1 central point) across four primary factors: position, velocity, acceleration, and direction of movement. To address the challenge of obscured lower body parts during practice sessions, we derived a central point by calculating the mean of the coordinates of the skeleton points situated above the nurse's waistline. This provided both the center point of the x and y coordinates. Position indices were obtained from the extracted skeleton data, while velocity was computed by determining the difference between consecutive coordinate values over time, as expressed in formula (1). Similarly, acceleration was calculated as the difference between consecutive velocity values over time, as shown in formula (2). The time interval between two consecutive values was fixed at 1/30 second. To determine the direction of movement, we employed boolean metrics, where the value equals 1 if moving forward and 0 if moving backward. This determination was based on the velocity value, as described in formula (3).

$$v = \Delta x \div 1/30 \tag{1}$$

$$a = \Delta v \div 1/30 \tag{2}$$

$$moving\_forward = \begin{cases} 1 & \text{if } v > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

After receiving coordinate, velocity, acceleration, and moving direction sequences for 14 body marks in both the x and y dimensions, statistical features were calculated to generate the final set of features. Initially, the data was scaled using a Robust scaler algorithm. Subsequently, 20 statistical-based features were extracted for every 3-second window segment. These features included mean, standard deviation, maximum, minimum, variance, median, sum, kurtosis, skewness, 25th and 75th percentiles, root mean square, mean absolute value, mode, root mean square error, mean absolute error, maximum error, minimum error, standard error, 25th percentile error, and 75th percentile error. Consequently, a total of 2240 features were extracted in this group of features.

b) Relative position and movement features: Due to variations in recording frames and the subject's relative position to the camera during data gathering, recorded positions may exhibit variability, resulting in unnormalized features extracted from these coordinates. To address this issue, relative coordinates were utilized in this study. The right shoulder point served as the reference point, establishing a system of vectors from this reference to the other 12 body marks in both the x and y axes. Subsequently, relative position features were redetermined from these vectors. Additionally, relative velocity, acceleration, and moving direction were computed based on relative position using formulas (1), (2), and (3), respectively.

Similar to the extraction of location and movement features, we computed a set of 20 statistical features for 3-minute segments of relative location, velocity, acceleration, and relative moving direction sequences. This process was performed in 2 axes of 12 relative body marks and yielded a total of 1920 features within this category.

c) Angels-based features: Joint angles play a crucial role in reflecting underlying activity by capturing pose and relative movement between body parts. The inclusion of this additional information enhances the model's capacity to interpret data accurately and make predictions. In this study, each joint angle is calculated using three marks through the vector formed by these coordinates (refer to Equation (4)).

$$Angle = arccos(\frac{\overrightarrow{\mathrm{AB}}.\overrightarrow{\mathrm{AC}}}{||\overrightarrow{\mathrm{AB}}|| \times ||\overrightarrow{\mathrm{AC}}||}) \tag{4}$$

where A is the coordinates of the middle point, the B and C are the coordinates of the remaining points.

In this study, a total of 9 angles were determined, including left elbow-shoulder-hip, right elbow-shoulder-hip, left wrist-elbow-shoulder, right wrist-elbow-shoulder, right elbow-right shoulder-left shoulder, left elbow-left shoulder-right shoulder, elbow-center-elbow, shoulder-center-shoulder, and wrist-center-wrist. Finally, similar to the extraction process for location and movement features, a total of 20 statistical metrics were computed for each segment from the determined angle sequences. Consequently, a total of 180 features were obtained in this group of features.

d) Time-lag features: Time series components demonstrate serial dependence, indicating that the value at a particular time point can be deduced from preceding time points [8].

, which involves utilizing past values of the target series as features, thus aligning them with the contemporaneous values we aim to predict (i.e., within the same temporal context). Specifically, by shifting feature values from the preceding 3 seconds of each point in the feature sequences, we augment our models with additional information to improve the classification performance. Moreover, time-lag features facilitate the identification of patterns and trends in time series data, while also imparting autoregressive characteristics to the chosen model, thereby simulating the utilization of past values for predicting future outcomes.

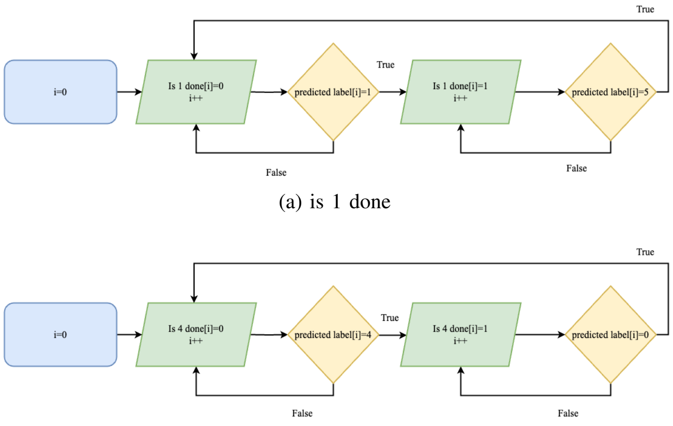
In this study, time-lag features were incorporated into three groups of features: 2240 position and movement features, 1920 relative position and movement features, and 180 angle-based features. Remarkably, these feature sets were reused from the immediately preceding segment.

Furthermore, the time feature, which represents the temporal aspect of the practice, was also included. This feature describes the timestamp at which this segment was executed and is defined in Equation (5).

$$time_i = real\_time_i / real\_time_{max} \tag{5}$$

where $real\_time_i$ is the time of record and $real\_time_{max}$ is the total time of the practice

e) Sequential-based features: In the context of medical specialty activities, a multitude of tasks are executed within specific procedural frameworks. Leveraging the completion status of these activities, it becomes possible to predict the probability of subsequent actions. In our study, we introduce two sequential features represented as boolean values: 'is 1 done' (corresponding to activity ID 1, 'Temporal removal of an artificial airway') and 'is 4 done' (associated with activity ID 4, 'Catheter disinfection'). These sequential features have the potential to mitigate the ambiguity encountered by learning models based on skeleton points when distinguishing between similar pose activities.



(a) is 1 done

(b) is 4 done

Fig. 1. The flow chart shows how the sequential-based feature was taken at frame i, the process will end at the end of the practice

- The "is-1-done" indice aims to distinguish 2 similar poses: activity ID 0 (Catheter preparation) and 4 (Catheter disinfection). According to the ES procedure, the catheter is prepared before and disinfection after removing the artificial airway from the patient.
- The "is-4-done" indice aims to distinguish invasion activity executed on the patient (1, 2,3) and post-ES activity (5,6,7).

In our study, we derive sequential features based on pre-classification results from four distinct feature groups. Initially, we set two sequential features to a value of 0, indicating that the corresponding activities have not yet been performed. Subsequently, when the predicted label aligns with the activity, these features transition to a value of 1. However, upon completion of the subsequent mandatory action, they are reset to 0. Specifically, we introduce two boolean features: 'is-1-done' (as shown in Figure 1a) and 'is-4-done' (as depicted in Figure 1b). These features serve as reference points for our model, facilitating optimal performance in distinguishing between activities.

D. Feature selection

In this study, to sift through and pinpoint the most pertinent and informative characteristics from the vast array of extracted features, the ANOVA F-test score was employed for feature importance calculation and selection. Specifically, the ANOVA F-test score was applied to 8683 extracted features across 5 groups. Subsequently, the time feature and 1999 features with the highest F-test scores were chosen for the classification phase. This procedure not only streamlines the data dimensionality but also accelerates model computation, diminishes resource consumption by excluding less meaningful features, and mitigates the risk of overfitting in classification.

E. Data augmentation by generative model

In order to enhance the performance of the 9-class classification, a data augmentation process was conducted by a generative model. This process aimed to increase the number

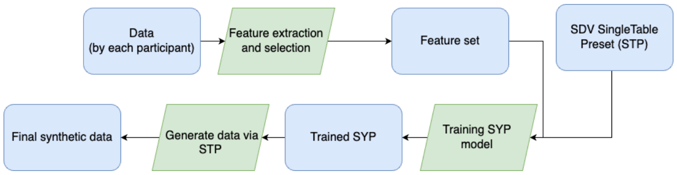


Fig. 2. The flow chart of procedure to data augmentation.

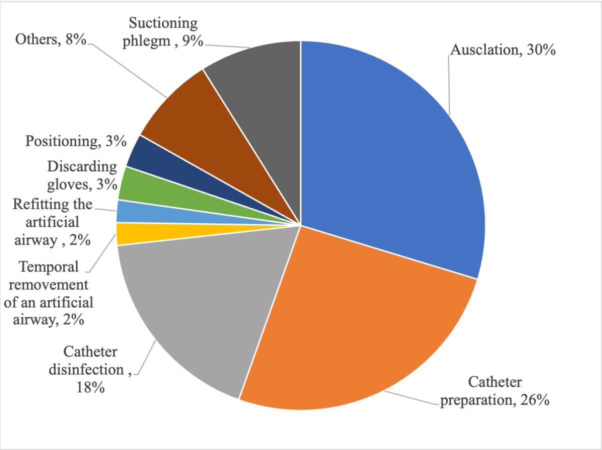


Fig. 3. The pie chart of distribution of training data by activity ID

of samples, particularly in the minority classes, based on tabular data of extracted features. In this work, the Synthetic Data Vault with Gaussian Copula (SDV-G) [9] was employed in this work to generate a single table of skeleton-based features.

SDV-G is adept at mimicking real-world datasets and is particularly effective with large tabular datasets [10]. This process involves computing statistics across interconnected database tables and employing advanced multivariate modeling techniques. Through an iterative approach, SDV-G constructs a comprehensive model of the entire database, which prevent overfit synthesis data and boost the generated data quality. Once the model is established, it allows for seamless data synthesis, enabling sampling from any database segment. As a result, SDV-G provides unparalleled versatility and utility in generating synthetic datasets for various analytical purposes [9].

In this study, SDV-G was applied to each participant's data, as illustrated in Figure 2. Initially, the top 2000 features based on ANOVA F_score were extracted from the original skeleton dataset. Subsequently, these features were inputted into the SDV-G model to train the data synthesis model. Following training, the SDV-G model was utilized to generate synthetic data, with the number of generated samples equal to that of the original dataset. Finally, time feature extraction is conducted on the generated data to produce the final augmented dataset.

F. Handling imbalance data

In this study, an oversampling technique was employed to enhance the performance of the classification model and address the imbalance in class distribution. As illustrated in Figure 3, the uneven distribution of data across classes is evident, with classes 0, 4, and 7 collectively constituting over 60% of the training data, indicating a significant class imbalance. To mitigate this issue and ensure fair representation across all classes, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized [11]. SMOTE effectively balances the training dataset by generating synthetic samples for minority classes, including classes 1, 2, 3, 5, 6, and 8. This approach helps to alleviate the class imbalance problem and promotes more accurate and reliable model predictions.

G. Classification model

To perform the classification of ES activity based on extracted and selected features, three ensemble learning models, Extreme Gradient Boosting Classifier (XGBoost), and Hist Gradient Boosting Classifier (HGBC), were employed in this work.

- Extreme Gradient Boosting Classification (XGBoost) is a boosting ensemble learning model that enhances classification performance by sequentially training models to correct the errors of previous models. XGBoost has gained popularity for its effectiveness in various machine learning tasks due to its ability to handle complex datasets and produce highly accurate predictions [12]. In this study, the XGBoost is set with evaluation metrics as mlogloss.
- Hist Gradient Boosting Classifier (HGBC): Similar to the XGBoost model, the HGBC model operates within the gradient boosting framework. It involves the sequential construction of decision tree models, with each tree aiming to rectify the errors of its predecessors. However, HGBC distinguishes itself by reducing training complexity through the binning of features into histograms. This summarization of data distribution is particularly advantageous for handling large feature datasets. In this study, HGBC is configured with balanced class weights, as described in the literature [13].
- Voting model: is a method that leverages multiple machine learning models to collectively decide on the final classification result [14] . In this work we use soft voting model, which is a combination of for trained machine learning model (two HGBC and two XGB) to vote and return the argument of maxima of the class have highest mean possibility to classify the activity.

In this study, we employed the three models described earlier in various combinations and datasets to achieve the best classification performance. By exploring different scenarios of model combination and dataset utilization, we aimed to optimize the classification accuracy and robustness of our system.

H. Model training and validation

The data set is divided depending on the individual into a training set and a validation set as follows:

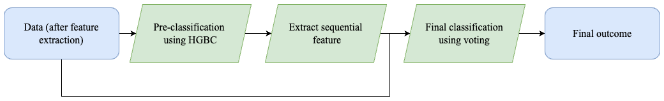

Data (after feature extraction)
Pre-classification using HGBC
Extract sequential feature
Final classification using voting
Final outcome

Fig. 4. The flowchart of classification pipeline

- The training set contains 4 nurses (ID 1, 4, 7, 12) and 6 student (ID 1, 5, 7, 9, 10, 11)
- The validation set contains 3 nurses (ID 2, 6, 11) and 3 students (ID 2, 3, 8).

To obtain sequential data, assumptions regarding activity labels are necessary. In this study, the HGBC model (pre-classify model) was utilized for pre-training on the original feature set following oversampling by SMOTE. The pre-defined labels were then employed to calculate two sequential features.

To generate the final predictions, both the original and synthesized feature sets were utilized in the primary training process across two classification models. Following the over-sampling of the minority class by SMOTE, the data was input into the models under two scenarios: using only the original data and incorporating both the original and synthesized data. Subsequently, four trained models were obtained: XGBoost (XGB1, XGB2), and Histogram-Based Gradient Boosting Classifier (HGBC1, HGBC2)[1]. Finally, the classification performance was evaluated across three scenarios:

- Voting model: Utilize the soft voting model of XGB1, XGB2, HGBC1, and HGBC2 to predict final prediction
- XGB+SDV: Utilizing the trained XGB1 model to predict the final prediction.
- XGB: Employing the trained XGB2 model to predict the final prediction.

I. Nurse and student activity analysis

In order to elucidate the distinctions in performance between activities conducted by experienced nurses and students, two distinct experiments were designed and executed:

- The first experiment aimed to assess the ES activity classification independently for experienced nurses and students. This approach was intended to examine the intra-variance of activity execution among practitioners.
- The second experiment focused on the classification of nurses and students within individual ES activities. The outcomes of this experiment are anticipated to provide insights into the extent to which an activity can reflect the practitioner's level of experience. Put differently, activities that yield high discrimination performance may suggest a greater need for intensified training for student practitioners.

Both experiments were performed using the proposed pipeline integrated with the XGB model.

TABLE II
THE OVERALL CLASSIFICATION RESULTS OF THE MODELS AND NGO ET AL. MODEL (IN PERCENT)

	Voting	**XGB + SDV**	**XGB**	**pre-classify**	**Ngo et al. model [2]**
Weighted avg F1-score mean	**89.9 ± 7.1**	86.0 ± 8.6	**89.9 ± 7.0**	86.8 ± 5.8	46.0±20.3
accuracy mean	**90.4 ± 7.3**	86.8 ± 8.5	90.3 ± 7.0	87.7 ± 5.9	53.8±15.0
weighted avg F1-score median	**92.5**	89.8	91.9	86.7	38.5
accuracy median	**93.9**	90.9	92.3	87.3	50.0

TABLE III
THE CLASSIFICATION RESULTS FOLLOW BY PARTICIPANT ID (IN PERCENT)

	Weighted avg F1 score			**Accuracy**		
	Voting	**XGB + SDV**	**XGB**	**Voting**	**XGB + SDV**	**XGB**
S08T1	90	88	92	90	89	92
S08T2	72	92	73	73	93	74
N11T1	94	95	95	96	97	96
N11T2	94	93	94	96	95	96
N06T1	85	69	79	84	66	79
N06T2	90	89	92	92	90	93
S02T2	88	85	89	89	86	89
S02T1	89	80	90	88	81	89
N02T2	92	87	93	94	89	94
N02T1	95	87	96	96	89	97
S03T1	95	77	94	95	77	94
S03T2	92	92	92	93	92	92

IV. RESULT AND ANALYSIS

A. Classification Results

The classification results, as shown in Tables II and III, demonstrate notable ES activity classification performance across three learning model scenarios. The Voting model, in particular, achieved the highest performance metrics, boasting a weighted average F1-score and accuracy of 89.9% and 90.4% respectively, which is an improvement of 44% and 36.6% to Ngo et al. research . Moreover, the Voting model marginally outperformed the XGB model in terms of median accuracy and weighted average F1-score, with improvements of approximately 1.5% and 0.6%. However, both the Voting model and XGB displayed slightly superior accuracy and F1 scores, approximately 4% higher, compared to the XGB + SDV model. This discrepancy may indicate potential over-fitting when applied to extensive datasets or imperfections in the synthetic records generated. These findings underscore the significance of ensemble modeling in achieving superior classification performance. Personal evaluations detailed in Table III further corroborate this, with the Voting classifier yielding impressive weighted F1-scores and accuracy. Notably, in certain participants (N06T1 and S03T1), the XGB model's performance on both original and synthetic data was modestly reduced, reflected in F1-scores of 69% and 77%.

A substantial number of extracted features were evaluated using ANOVA F-test scores, as depicted in Figure 5. Among

[1] *1 model denotes the model fitted using the combination of original and SDV-generated feature sets, while *2 model indicates the model fitted with the original data only.

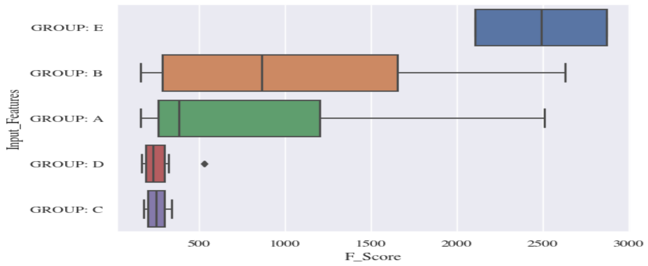


Fig. 5. The boxplot of F-scores across 5 groups of features

TABLE IV
THE CLASSIFICATION PERFORMANCE OF ES ACTIVITY BY THE XGB MODEL OF NURSE AND STUDENT (IN PERCENT)

	Student	Nurse
Weighted avg F1-score mean	88± 8	92± 4
Weighted avg F1-score median	89	93
Accuracy mean	88	93
Accuracy median	89	95

these, the proposed sequential features (Group E) exhibited the highest average F-score, underscoring the significance of activity order. Both the raw (Group A) and relative (Group B) location and movement features also demonstrated high F-scores, with interquartile ranges approximately from 250 to 1250 and 1750, respectively. Notably, the relative location and movement features achieved marginally higher scores due to the correction of recording frame variance. Meanwhile, the time-lag and angle-based features yielded F-scores ranging from 100 to 400.

B. Nurse and Student Performance Comparison

It is noteworthy that, as indicated in Table IV, the model demonstrates superior classification performance for nurse activities. Specifically, the mean weighted average F1 score for nurse activity recognition stands at 92%, which is 5% higher than that of the students. Additionally, the standard deviation of the weighted F1 score for students is greater than that for nurses, with values of 8% and 4%, respectively. This disparity may be attributed to the nurses' experience and precision in performing actions compared to the students.

In the classification of experienced nurses and students, discernible differences in discriminative performance were observed across various activities, as detailed in Table V. Basic activities such as Catheter preparation, Catheter disinfection, Discarding gloves, and Auscultation yielded lower accuracy rates (63%, 59%, and 66% respectively), indicating a significant overlap in the performance of these activities between experienced nurses and students. Conversely, more complex activities that are directly performed on patients and require a well-trained process, such as Temporal removal of an artificial airway, Suctioning phlegm, Refitting the artificial airway, and Positioning, demonstrated high discriminative performance with accuracy rates of 88%, 80%, 82%, and 80% respectively.

TABLE V
NURSE AND STUDENT CLASSIFICATION PERFORMANCE BY ACTIVITY ID

Activity ID	Name of activity	Accuracy
0	Catheter preparation	63
1	Temporal removement of an artificial airway	88
2	Suctioning phlegm	80
3	Refitting the artificial airway	82
4	Catheter disinfection	59
5	Discarding gloves	68
6	Positioning	80
7	Auscultation	66
8	Others	71

TABLE VI
MEAN F1-SCORE (IN PERCENT) OF THE CLASSIFICATION MODEL WITH AND WITHOUT SDV BY ACTIVITY ID

	Voting with SDV data[2]	Voting without SDV data[3]
Activity ID 1	74	66
Activity ID 2	91	90
Activity ID 3	62	68
Activity ID 5	69	67
Activity ID 6	69	71
Activity ID 8	89	90
Average	76	75

V. DISCUSSION

This study has adeptly identified a robust set of features and established an effective learning pipeline that exhibits high performance in ES activity classification. Notably, the voting model, which integrates XGB and HGBC models trained on both original and synthesized feature sets, demonstrated superior performance. This outcome validates the efficacy of ensemble learning when dealing with datasets that contain a vast number of features and highlights the utility of synthesized data in addressing class imbalance and the challenges posed by multi-class classification tasks by boost the the average F1 score of minority class by about 1%, and the F1-score in activity ID 1 classification by 8% (Table VI). However, it was observed that the synthesized feature set was not optimally generated, leading to a diminished performance of the XGB model, as indicated in Table II. Consequently, there is a potential avenue for exploration in more robust tabular generative AI models, such as TABGAN, Variational Autoencoders (VA), Normalizing Flows (NF), diffusion technique (DFM), advancements in image data generation techniques, and the combination of copula function with above mention generative models such as CopulaGAN.

In terms of feature impact, the location and movement features (Groups A and B) significantly influenced model training, effectively capturing the variance in poses across different ES activities, particularly in relation to the right shoulder reference point. Furthermore, the two newly proposed sequential features made a substantial contribution to the model training process, aiding in the differentiation of similar poses within medical procedures by their sequence.

[2] A voting classifier consist of HGBC1, XGB1, HGBC2, and XGB2

[3] A voting classifier consist of HGBC2, and XGB2

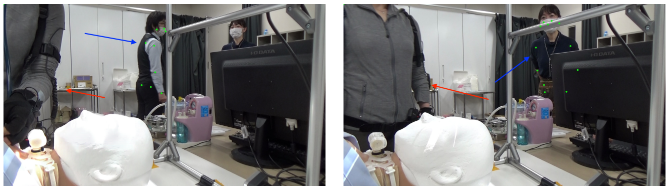


Fig. 6. The YOLOv7 catching wrong person (blue arrow indicate the person YOLOv7 selected, and the red arrow indicate the practitioner)

This study has observed a distinct gap in the performance of activities by experienced nurses compared to students, particularly in tasks that demand a high level of expertise. According to the classification results presented in Table V, there is a pronounced disparity in the execution of complex activities such as Temporal removal of an artificial airway, Suctioning phlegm, and Refitting the artificial airway. These activities require advanced skills due to their invasive nature, leading to a marked distinction in performance accuracy between nurses and students. Conversely, for non-invasive activities like the preparation and disinfection of catheters, no significant differences were noted between the two practitioner groups. This finding suggests that ES activity recognition systems could be employed to assess the proficiency level of practitioners or to provide corrective feedback for novices, thereby offering substantial value in nursing education.

During our observations, we noted errors in the extraction of skeleton points when multiple individuals appeared within a video frame. Specifically, in the case of N06T1, the predictive analysis revealed that at approximately 4 minutes and 17 seconds, the keypoints identified by YOLOv7 erroneously targeted an incorrect individual—a man behind the screen followed by a woman adjacent to him—rather than the practitioner, as illustrated in Fig. 6. To rectify this issue, we propose a method that selects keypoints from the individual whose cumulative distance from the origin (0,0) of the video frame and the most recently recorded right shoulder coordinate to the current right shoulder point is minimal.

To augment the performance of behavior classification, an integrative approach that combines skeleton data with textual descriptions could be considered. Rather than relying exclusively on skeleton data, this method introduces supplementary action descriptions. These descriptions can be synthesized by Large Language Models (LLMs) like GPT-3 or crafted manually by domain experts [15], [16]. In the manual approach, experts annotate key body movements over time for each action, enriching the existing dataset. The synergy of skeletal and textual data offers a holistic view of behavior. While the former provides quantitative insights into motion and body positioning, the latter qualitatively details the actions' intent and execution. This dual-faceted analysis fosters a comprehensive contextual understanding, thereby bolstering the predictive prowess of behavior classification models.

VI. Conclusion

This research delineated novel feature extraction strategies and a model training pipeline for the classification of nursing activities during endotracheal suctioning (ES).

The study's cornerstone is the deployment of SDV-G data generator and an ensemble learning model that amalgamates two XGBoost and two HGBC models (soft voting classifier), which has demonstrated commendable efficacy in discerning varied activities inherent to the ES procedure by achieving the high accuracy of 90.4%. These results are instrumental in advancing the automated recognition of ES activities. Additionally, the research shed light on the comparative analysis between nurses with three years of experience and nursing students. This comparison has yielded insights that are pivotal in formulating recommendations for enhancing nursing training programs.

References

[1] H. A. V. Ngo, Kaneko H., Hassan I., Ronando E., Shoumi M., Munemoto R., Hossain T. & Inoue, "S. Summary of the Nurse Care Activity Recognition Challenge Using Skelton Data from Video with Generative AI", *International Journal Of Activity And Behavior Computing*, vol. 2024, 2024.

[2] H. A. V. Ngo, Q. N. P. Vu, N. Colley, S. Ninomiya, S. Kanai, S. Komizunai, A. Konno, M. Nakamura, S. Inoue, "Toward Recognizing Nursing Activity in Endotracheal Suctioning Using Video-based Pose Estimation", *International Journal of Activity and Behavior Computing*, vol. 2024, no. 1, pp. 1, 2024, doi: 10.60401/ijabc.1.

[3] S. Komizunai et al., "An interactive endotracheal suctioning simulator which exhibits vital reactions: ESTE-SIM," *International Journal of Automation Technology*, vol. 13, no. 4, pp. 490–498, Jul. 2019, doi: 10.20965/ijat.2019.p0490.

[4] A. A. Halim et. al, "Human action recognition based on 3D skeleton part-based pose estimation and temporal multi-resolution analysis," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016, pp. 3041-3045, doi: 10.1109/ICIP.2016.7532918.

[5] F. Meng, H. Liu, Y. Liang, J. Tu, and M. Liu, "Sample Fusion Network: an End-to-End data augmentation network for Skeleton-Based human action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5281–5295, Nov. 2019, doi: 10.1109/tip.2019.2913544.

[6] H. Ramirez, S. A. Velastín, S. Cuéllar, E. Fábregas, and G. Farías, "BERT for Activity Recognition Using Sequences of Skeleton Features and Data Augmentation with GAN," *Sensors*, vol. 23, no. 3, p. 1400, Jan. 2023, doi: 10.3390/s23031400.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *arXiv:2207.02696 [cs.CV]*, Jun. 2023.

[8] O. Surakhi et al., "Time-Lag selection for Time-Series forecasting using neural network and heuristic algorithm," *Electronics*, vol. 10, no. 20, p. 2518, Oct. 2021, doi: 10.3390/electronics10202518.

[9] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," *2016 IEEE international conference on data science and advanced analytics (DSAA)*, 2016, doi: 10.1109/dsaa.2016.49.

[10] M. Endres, A. M. Venugopal, and T. S. Tran, "Synthetic Data Generation: A Comparative Study," *IDEAS '22: Proceedings of the 26th International Database Engineered Applications Symposium*, 2022. doi: 10.1145/3548785.3548793.

[11] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem: A Review," *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1-8, doi: 10.1109/ICIC54025.2021.9632912.

[12] XGBoost Documentation. (2022). Accessed: Feb 3, 2024.[Online]. Available: https://xgboost.readthedocs.io/en/stable/#xgboost-documentation

[13] G. Greffenstette, M. Kampert, and P. Kranen, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *2019 International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6017-6023, 2019.

[14] Y. Zhang, H. Zhang, J. Cai, and B. Yang, “A weighted voting classifier based on differential evolution,” *Abstract and Applied Analysis*, vol. 2014, pp. 1–6, Jan. 2014, doi: 10.1155/2014/376950.

[15] W. Xiang, C. Li, Y. Zhou, B. Wang and L. Zhang, ”Generative Action Description Prompts for Skeleton-based Action Recognition,” *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 10242-10251, doi: 10.1109/ICCV51070.2023.00943.

[16] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, “Language knowledge-assisted representation learning for skeleton-based action recognition,” *arXiv:2305.12398 [cs.CV]*, May. 2023.