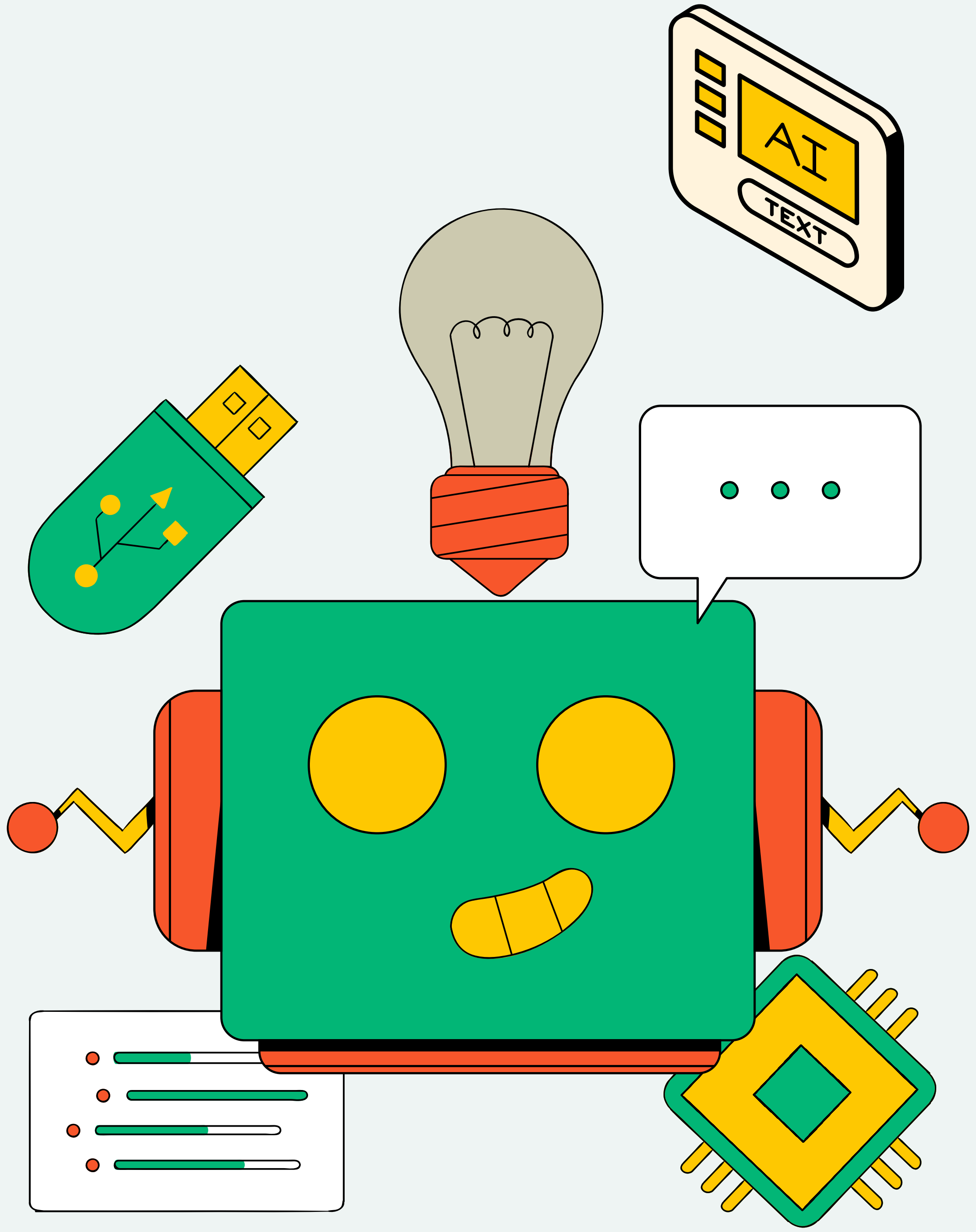
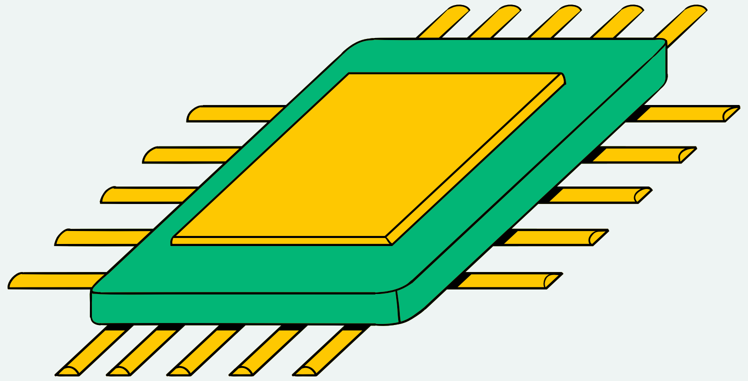


Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity

Hyewon Han & Junhyoung Oh

Report by James Sablay



CMSC 197 | October 2, 2024

Obstructive Sleep Apnea Syndrome

- Obstructive sleep apnea (OSAS) affects **1 billion adults globally**, raising **cardiovascular and cognitive risks.**

- This oftens remain **undiagnosed**, highlighting the need for a more active treatment.

Common tools for diagnosis

- **Polysomnography (PSG)** is the gold standard for diagnosing OSAS, but its complexity and use of sensors can disrupt sleep.
- With the rise in suspected OSAS cases, there is a growing need for **simpler diagnostic methods.**

Use of machine learning in OSAS

- Machine learning is increasingly applied in healthcare for its ability to **handle complex data.**
- In OSAS research, clustering **helps phenotyping** and improves classification, making it a key tool for **better prediction models.**
- **Develop** machine learning models to predict OSAS severity **without requiring PSG.**

Data acquisition

Data Source: Patients from Samsung Medical Center's sleep clinic (2014-2021)

Data Collected:

- **Personal Information:** Gender, age, height, weight
- **Physical Measurements:** Abdominal, neck, and hip circumference
- **Self-Reported Questionnaires:** Epworth Sleepiness Scale (ESS), Insomnia Severity Index (ISI)
- **PSG Method:** Conducted using Embla N7000 with automated scoring to determine OSAS and measure apnea-hypopnea index (AHI).

Data pre-processing

Dataset Overview:

- **Samples:** 4,014
- **Features:** 33 numerical and categorical characteristics
- **OSAS Severity Classification:** Four classes based on American Academy of Sleep Medicine guidelines

Data pre-processing

- **Test Data:** 20% of the dataset
- **Classifier Training:** 5-fold cross-validation on training data
- **Statistical Analysis:**
    - Normal distribution assessed using the Kolmogorov-Smirnov test
    - Student's t-test for normally distributed data; Mann-Whitney U test for non-normal data
    - Chi-square test for categorical features
    - **Significance level:** $p < 0.05$

Table 1. The report of the statistical analysis.

Feature	All (n = 4014)	Training (n = 3211)	Test (n = 803)	*p*-value
Demographic parameters				
Age (years)	53.0 [40.0, 62.0]	53.0 [40.0, 62.0]	54.0 [40.0, 62.0]	0.282
Sex	Male: 2841 (70.78%),	Male: 2289 (71.29%)	Male: 552 (68.74%)	0.956
	Female: 1173 (29.22%)	Female: 922 (28.71%)	Female: 251 (31.26%)	
BMI (kg/m^2)	25.2 [23.0, 27.7]	25.1 [23.0, 27.7]	25.3 [22.9, 28.0]	0.387
Height (cm)	168.0 [161.0, 174.0]	168.0 [162.0, 174.0]	168.0 [161.0, 173.0]	0.383
Weight (kg)	71.0 [62.0, 80.0]	71.0 [62.0, 80.0]	71.0 [62.0, 80.0]	0.914
Body measurements				
Abdominal circumference (cm)	90.0 [83.0, 97.0]	90.0 [83.0, 97.0]	90.0 [83.0, 97.0]	0.704
Head circumference (cm)	57.0 [55.0, 58.0]	57.0 [55.0, 58.0]	57.0 [55.0, 58.0]	0.587
Hip circumference (cm)	96.0 [92.0, 100.0]	96.0 [92.0, 100.0]	96.0 [92.0, 100.0]	0.695
Neck circumference (lying position, cm)	38.5 [35.5, 40.5]	38.5 [35.5, 40.5]	38.5 [35.5, 40.5]	0.572
Neck circumference (sitting position, cm)	38.0 [35.0, 40.0]	38.0 [35.0, 40.0]	38.0 [35.0, 40.0]	0.661
Sleep questionnaires				
ESS scores	9.0 [6.0, 13.0]	9.0 [6.0, 13.0]	9.0 [6.0, 13.0]	0.308
ISI scores	11.0 [7.0, 16.0]	11.0 [7.0, 16.0]	12.0 [7.0, 16.0]	0.347
K-BDI-II scores	11.0 [7.0, 17.0]	11.0 [7.0, 17.0]	11.0 [7.0, 17.0]	0.765
PSQI scores	7.0 [5.0, 10.0]	7.0 [5.0, 10.0]	7.0 [5.0, 10.0]	0.636
SSS scores	3.0 [2.0, 3.0]	3.0 [2.0, 3.0]	3.0 [2.0, 3.0]	0.261
Other self-reported parameters				
Hours of sleep	6.0 [5.0, 7.0]	6.0 [5.0, 7.0]	6.0 [5.0, 7.0]	0.755
Consumption of hypnotics	Yes: 484 (12.06%)	Yes: 380 (11.83%)	Yes: 104 (12.95%)	0.973
	No: 3530 (87.94%)	No: 2831 (88.17%)	No: 699 (87.05%)	
PSG parameters				
AHI	20.4 [7.7, 40.4]	20.4 [7.85, 39.8]	20.1 [7.1, 42.45]	0.82
OSAS severity				
Normal	706 (17.59%)	565 (17.6%)	141 (17.56%)	0.261
Mild	897 (22.35%)	717 (21.33%)	180 (22.42%)	
Moderate	971 (24.19%)	777 (24.2%)	194 (24.16%)	
Severe	1440 (35.87%)	1152 (35.88%)	288 (35.87%)	

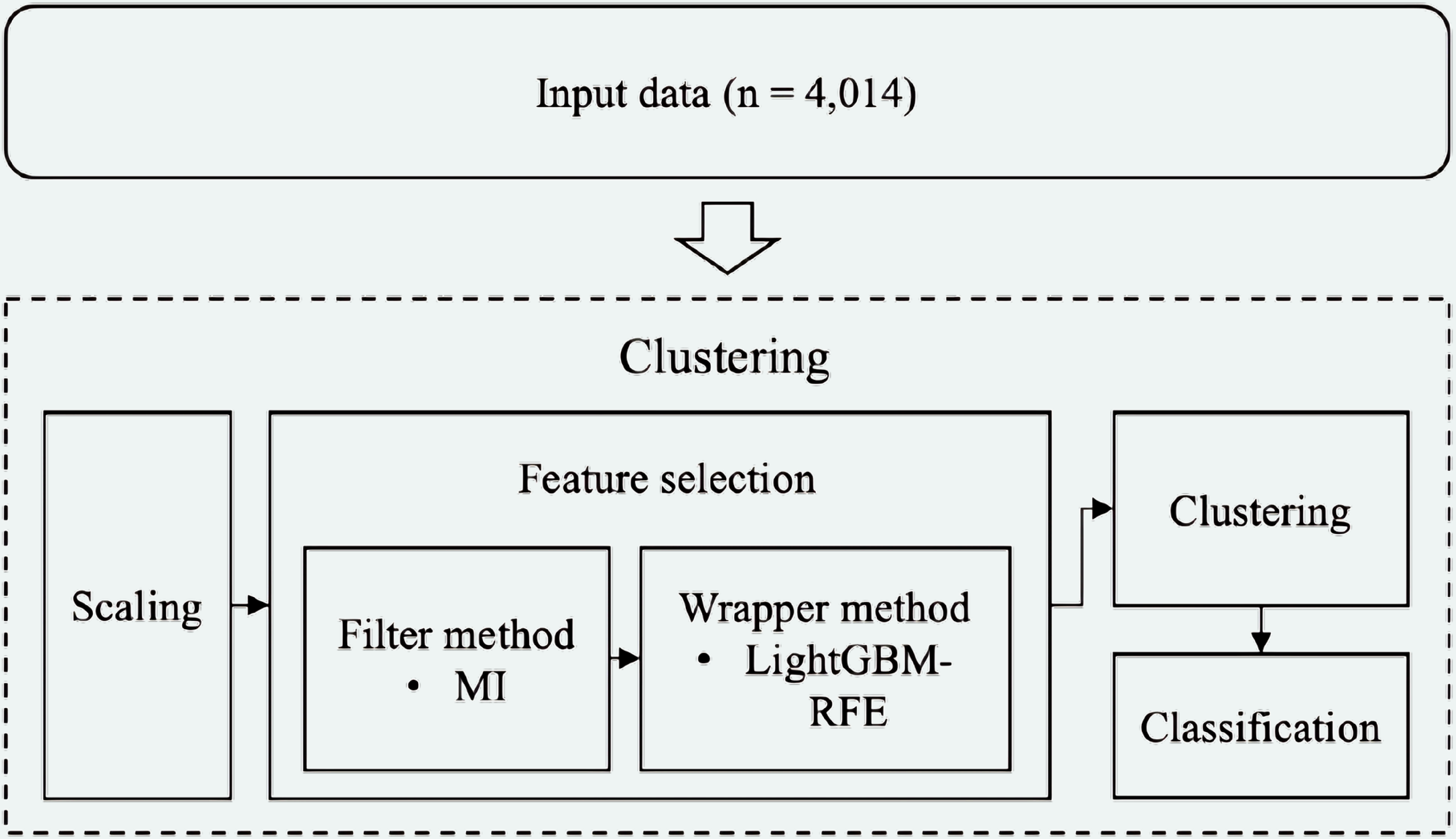


Figure 1. The workflow of the predictive models.

FEATURE SELECTION

- **Mutual Information (MI):** Measures interdependence between variables; used to filter less informative features with a threshold set at the mean MI score.
- **Recursive Feature Elimination (RFE):** Removes less important features iteratively to finalize the feature set for clustering.

CLUSTERING ALGOS

- Hierarchical Agglomerative Clustering
- K-means
- Bisecting K-means
- Gaussian Mixture Model (GMM)

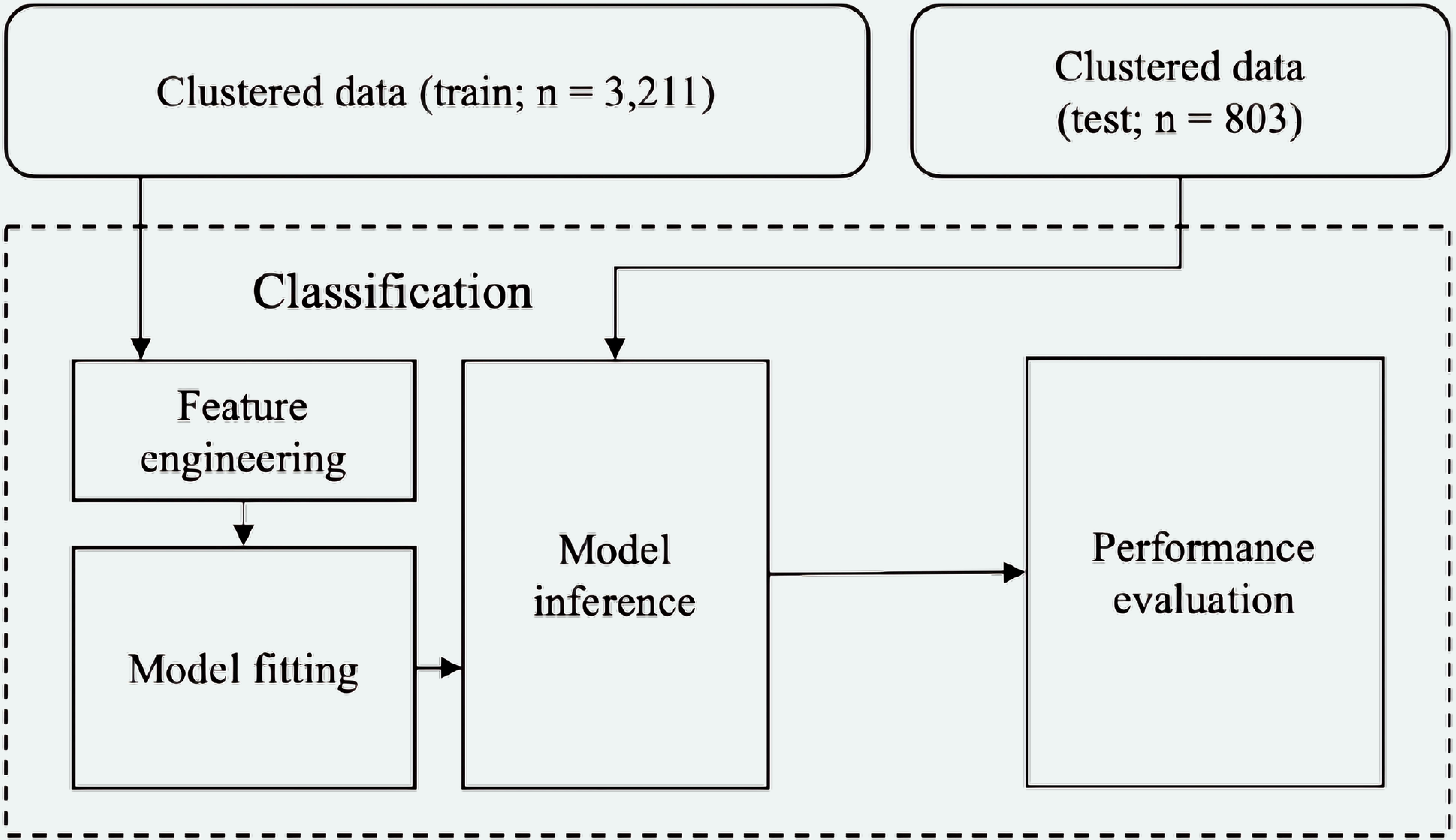


Figure 1. The workflow of the predictive models.

Feature engineering

Methods Used:

- **Weighted Epworth Sleepiness Scale (ESS):** Assigns different weights to each question, improving prediction accuracy for OSAS severity compared to the general ESS.
- **Predictive formula for AHI:**

$$Formula:\ AHI_{pred} = NC \times 0.84 + EDS \times 7.78 + BMI \times 0.91 - [8.2 \times gender\ constant(1\ or\ 2) + 37]$$

Modified using the SciPy package to optimize for the dataset, accounting for neck circumference measurements in both sitting and lying positions.

Feature engineering

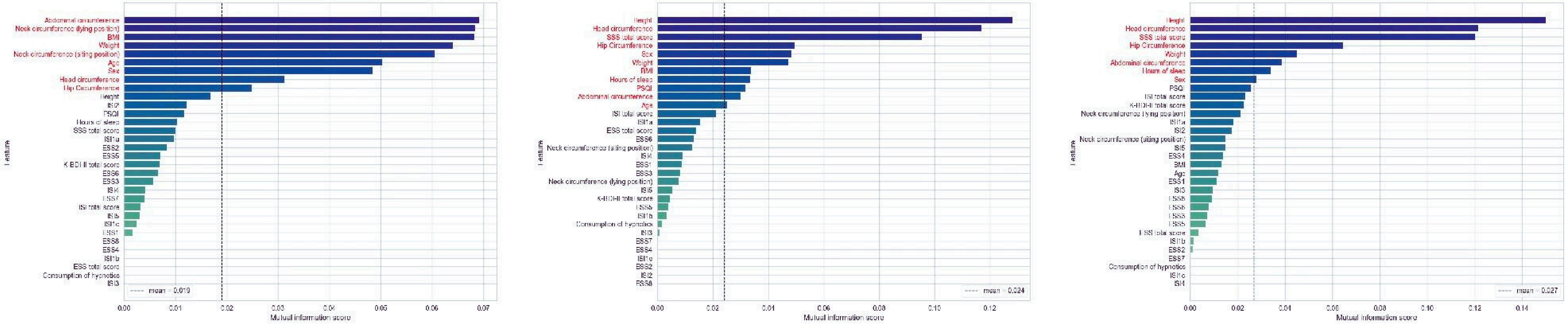
- **Additional Criteria:** Three criteria were utilized to assess excessive daytime sleepiness (EDS)
  - Weighted ESS criteria
  - Criteria from the American Academy of Sleep Medicine Task Force
  - Criteria from the study that proposed the predictive formula

Predictive models

Gradient boosting models and random forests (RF) are effective for large, complex datasets due to their accuracy and efficiency.

- **Models Used**
  - Random Forest (RF)
  - XGBoost
  - LightGBM
  - CatBoost

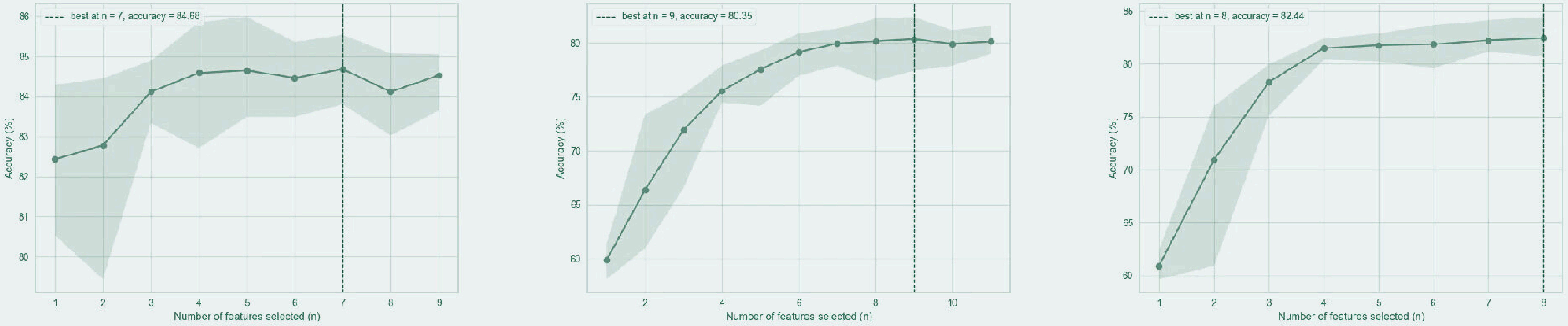
Clustering results



(a) AHI $\geq$ 5 (b) AHI $\geq$ 15 (c) AHI $\geq$ 30

Figure 2. Mutual information (MI) scores for all input features.

Clustering results



(a) AHI $\geq$ 5 (b) AHI $\geq$ 15 (c) AHI $\geq$ 30

Figure 3. Visualised 5-fold cross-validation results of recursive feature elimination (RFE).

Classification results by ML models and FE methods

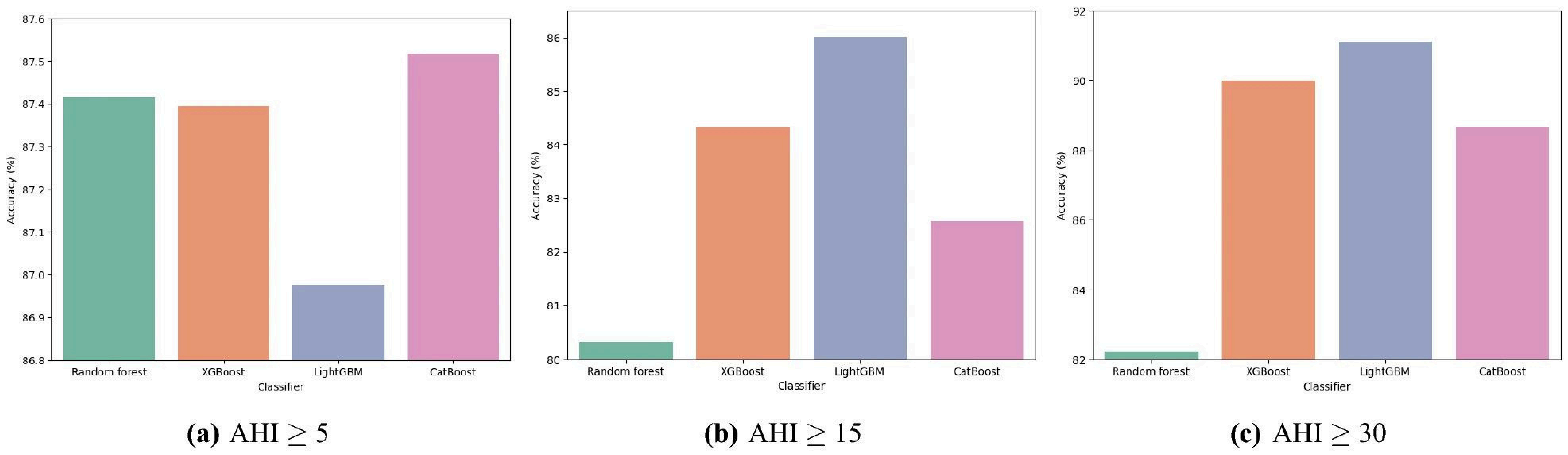


Figure 4. Comparisons of classification accuracy by machine learning classification algorithms.

Classification results by ML models and FE methods

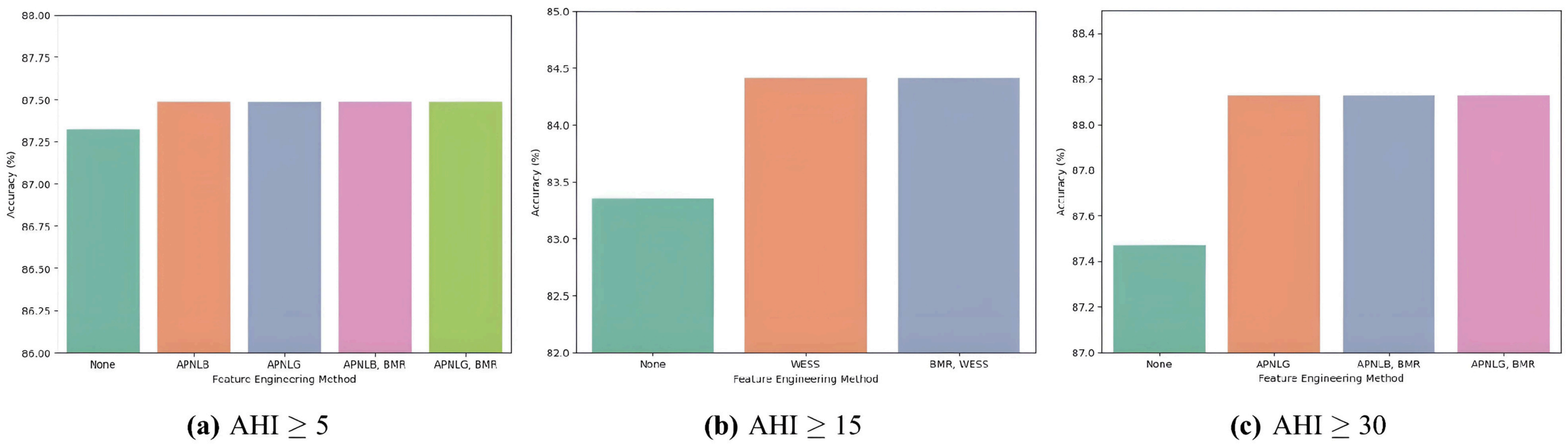


Figure 5. Comparisons of classification accuracy by feature engineering methods.

Predictive model building approach	AHI cut-off value	Accuracy (%)	AUC (%)	f1 (%)	Precision (%)	Recall (%)	*p*-value
Without clustering	5	86.18 (0.10) [86.05, 86.30]	84.83 (0.19) [84.55, 85.03]	91.98 (0.04) [91.93, 92.03]	88.14 (0.49) [87.36, 88.64]	96.19 (0.61) [95.47, 97.13]	**
	15	81.29 (1.57) [78.58, 82.44]	89.93 (1.32) [87.66, 90.85]	84.63 (1.18) [82.59, 85.48]	83.55 (1.70) [80.63, 84.87]	85.74 (0.65) [84.65, 86.31]	**
	30	89.13 (3.85) [82.57, 92.40]	95.17 (2.56) [90.80, 97.26]	83.99 (6.07) [73.58, 88.93]	88.32 (4.68) [80.58, 93.16]	80.12 (7.19) [67.71, 85.07]	**
Clustering only	5	88.14 (0.20) [87.76, 88.32]	82.31 (0.67) [81.29, 83.40]	93.06 (0.10) [92.90, 93.16]	89.11 (0.33) [88.56, 89.45]	97.40 (0.31) [96.87, 97.71]	$p < 0.05^*$
	15	85.75 (1.00) [84.00, 86.56]	92.89 (0.19) [92.70, 93.24]	89.03 (0.61) [87.96, 89.56]	86.43 (1.19) [84.38, 87.45]	91.86 (0.26) [91.33, 91.99]	$p < 0.05^*$
	30	90.74 (0.15) [90.49, 90.85]	95.09 (0.16) [94.89, 95.22]	78.53 (0.52) [77.54, 78.87]	88.42 (0.60) [87.45, 89.35]	71.67 (0.89) [69.89, 72.27]	0.635
Clustering with feature engineering	5	88.16 (0.25) [87.76, 88.42]	82.66 (1.24) [80.26, 83.74]	93.11 (0.15) [92.90, 93.28]	88.91 (0.32) [88.56, 89.45]	97.76 (0.47) [96.87, 98.07]	$p < 0.05^*$
	15	85.80 (1.51) [82.78, 86.63]	92.91 (0.24) [92.63, 93.36]	89.08 (0.92) [87.23, 89.60]	86.57 (1.67) [83.24, 87.47]	91.82 (0.19) [91.44, 91.99]	$p < 0.05^*$
	30	90.92 (0.20) [90.58, 91.12]	95.22 (0.20) [94.85, 95.39]	78.76 (0.49) [77.87, 79.20]	88.49 (0.79) [87.12, 89.58]	72.04 (0.91) [70.30, 72.89]	0.278
Clustering with feature engineering and hyperparameter tuning	5	87.82 (0.37) [87.30, 88.23]	81.56 (1.55) [79.56, 83.81]	92.95 (0.25) [92.65, 93.22]	88.56 (0.49) [88.05, 89.37]	97.85 (0.87) [96.18, 98.55]	$p < 0.05^*$
	15	87.84 (1.88) [85.53, 89.37]	95.02 (0.44) [94.37, 95.76]	90.79 (1.16) [89.37, 91.73]	89.27 (2.06) [86.75, 90.96]	92.65 (0.00) [92.65, 92.65]	$p < 0.05^*$
	30	91.06 (1.77) [87.63, 92.37]	95.03 (0.83) [93.43, 95.73]	75.66 (12.06) [51.62, 82.72]	90.87 (3.53) [86.87, 97.32]	70.19 (11.63) [47.06, 76.77]	0.056

Table 2. The report of classification metrics of predictive models by approaches.

RESULTS

Predictive model building approach	AHI cut-off value	Accuracy (%)	AUC (%)	f1 (%)	Precision (%)	Recall (%)	p-value
Without clustering	5	86.18 (0.10) [86.05, 86.30]	84.83 (0.19) [84.55, 85.03]	91.98 (0.04) [91.93, 92.03]	88.14 (0.49) [87.36, 88.64]	96.19 (0.61) [95.47, 97.13]	**
	15	81.29 (1.57) [78.58, 82.44]	89.93 (1.32) [87.66, 90.85]	84.63 (1.18) [82.59, 85.48]	83.55 (1.70) [80.63, 84.87]	85.74 (0.65) [84.65, 86.31]	**
	30	89.13 (3.85) [82.57, 92.40]	95.17 (2.56) [90.80, 97.26]	83.99 (6.07) [73.58, 88.93]	88.32 (4.68) [80.58, 93.16]	80.12 (7.19) [67.71, 85.07]	**
Clustering only	5	88.14 (0.20) [87.76, 88.32]	82.31 (0.67) [81.29, 83.40]	93.06 (0.10) [92.90, 93.16]	89.11 (0.33) [88.56, 89.45]	97.40 (0.31) [96.87, 97.71]	$p < 0.05$*
	15	85.75 (1.00) [84.00, 86.56]	92.89 (0.19) [92.70, 93.24]	89.03 (0.61) [87.96, 89.56]	86.43 (1.19) [84.38, 87.45]	91.86 (0.26) [91.33, 91.99]	$p < 0.05$*
	30	90.74 (0.15) [90.49, 90.85]	95.09 (0.16) [94.89, 95.22]	78.53 (0.52) [77.54, 78.87]	88.42 (0.60) [87.45, 89.35]	71.67 (0.89) [69.89, 72.27]	0.635
Clustering with feature engineering	5	88.16 (0.25) [87.76, 88.42]	82.66 (1.24) [80.26, 83.74]	93.11 (0.15) [92.90, 93.28]	88.91 (0.32) [88.56, 89.45]	97.76 (0.47) [96.87, 98.07]	$p < 0.05$*
	15	85.80 (1.51) [82.78, 86.63]	92.91 (0.24) [92.63, 93.36]	89.08 (0.92) [87.23, 89.60]	86.57 (1.67) [83.24, 87.47]	91.82 (0.19) [91.44, 91.99]	$p < 0.05$*
	30	90.92 (0.20) [90.58, 91.12]	95.22 (0.20) [94.85, 95.39]	78.76 (0.49) [77.87, 79.20]	88.49 (0.79) [87.12, 89.58]	72.04 (0.91) [70.30, 72.89]	0.278
Clustering with feature engineering and hyperparameter tuning	5	87.82 (0.37) [87.30, 88.23]	81.56 (1.55) [79.56, 83.81]	92.95 (0.25) [92.65, 93.22]	88.56 (0.49) [88.05, 89.37]	97.85 (0.87) [96.18, 98.55]	$p < 0.05$*
	15	87.84 (1.88) [85.53, 89.37]	95.02 (0.44) [94.37, 95.76]	90.79 (1.16) [89.37, 91.73]	89.27 (2.06) [86.75, 90.96]	92.65 (0.00) [92.65, 92.65]	$p < 0.05$*
	30	91.06 (1.77) [87.63, 92.37]	95.03 (0.83) [93.43, 95.73]	75.66 (12.06) [51.62, 82.72]	90.87 (3.53) [86.87, 97.32]	70.19 (11.63) [47.06, 76.77]	0.056

Table 2. The report of classification metrics of predictive models by approaches.

Discussion

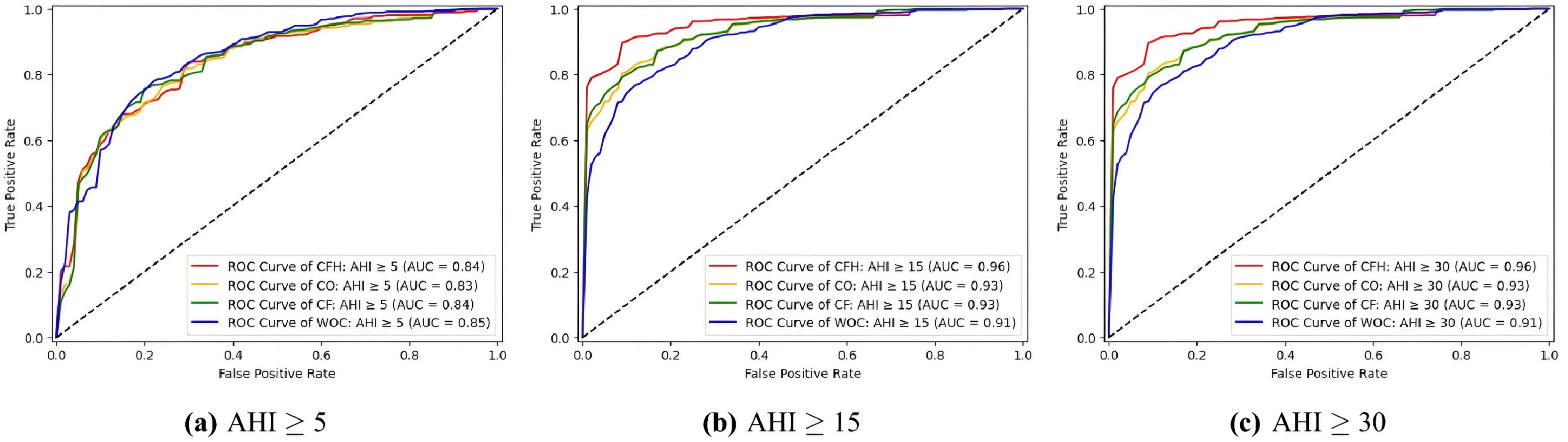
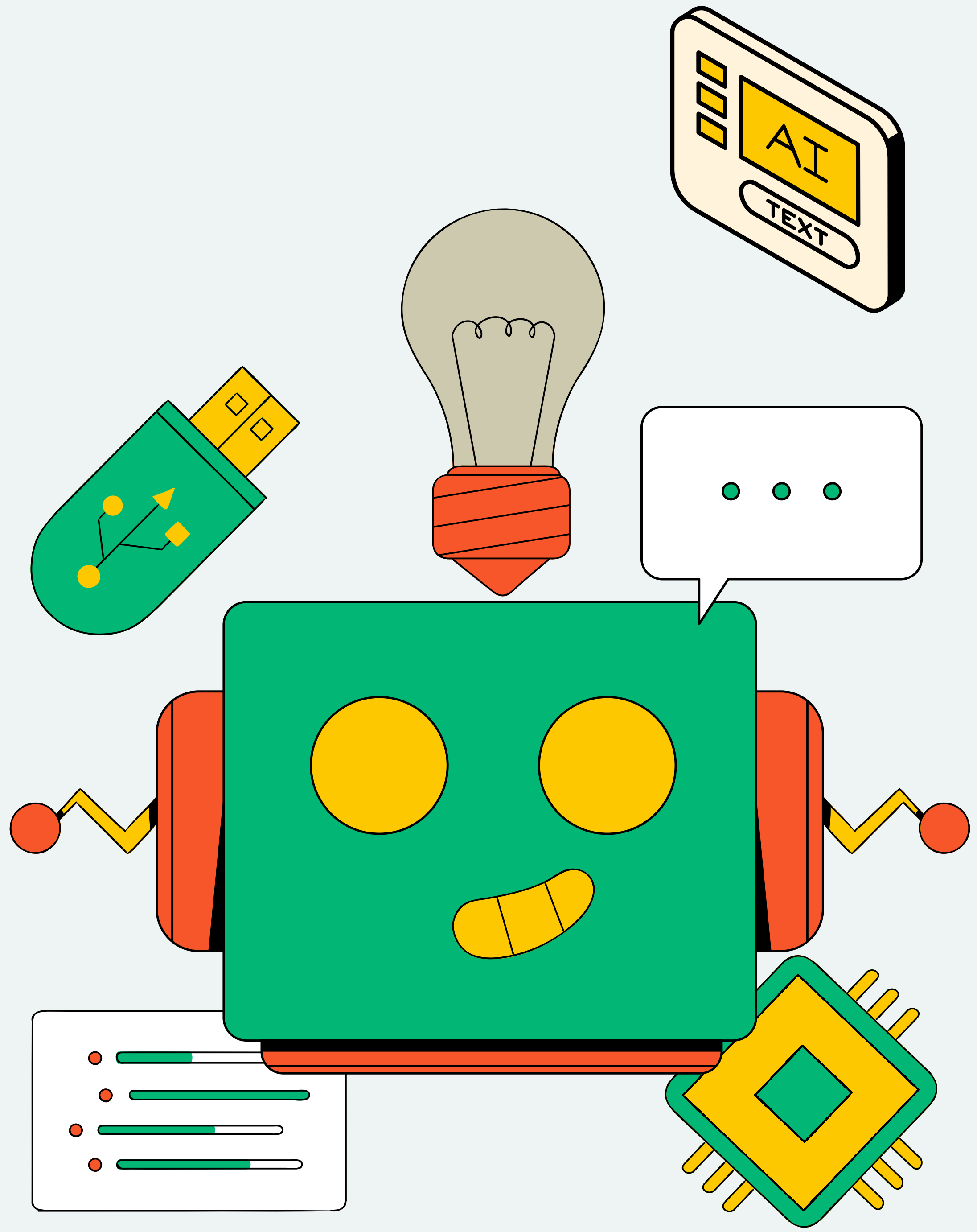


Figure 6. Comparisons of receiver operation characteristic(ROC) curves based on approach to building predictive models.

Conclusion

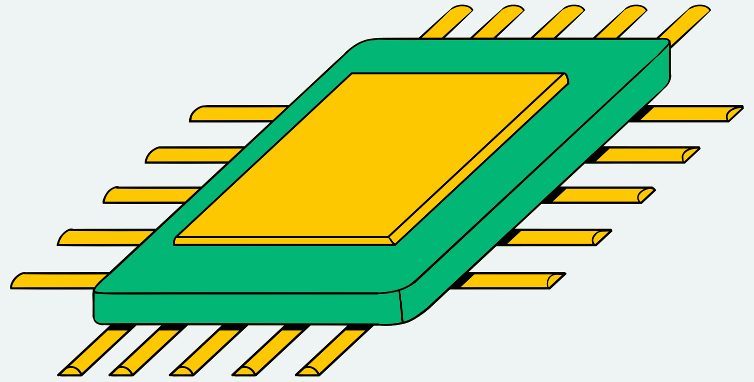
- Predicted **OSAS severity** using basic data (gender, age, body measurements, questionnaires) with **diverse machine-learning techniques.**
- Improved performance observed through a combination of **supervised and unsupervised learning**.
- **Limitations:** Missing data and single-institution focus; future research should include larger, multi-institutional datasets.
- The **predictive model** enables efficient OSAS screening.



Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity

Hyewon Han & Junhyoung Oh

Report by James Sablay



CMSC 197 | October 2, 2024