Supplementary Table 4. Results Based on Data Preparation in CRISP-DM (N = 125)

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A1	Marketta Hiissa et al. (2006)	Stemming	Word frequency, TF-IDF, N-grams, Stemming or Lemmatization results, Presence or absence of key terms	NI	Breathing, Blood Circulation, Pain	Train: 708, Test: 655
A2	Manabu Nii et al. (2007)	Morphological analysis, TF-IDF, Tokenization into morphemes	Number of morphemes, Number of terms in term list, Sum of TF-IDF weights, Specific medical/nursing terms presence	4	Nursing text classification (Classes 0 to 3: bad to very good)	10-fold cross-validation
A3	Laurence G. Moseley et al. (2008)	Categorization	Age, gender, entry qualifications, branch of nursing, grades awarded each semester, gross and net attendance records	6	Student dropout	NI
A4	Alexander Zlotnik et al. (2015)	Transformation, Standardized, Removal of irrelevant data	Day of week, holiday (yes/no), month of the year, week number	4	The number of ED visits	Train, Validation, Test: 2008-2012
A5	Manabu Nii et al. (2016)	Morphological analysis, Tokenization into morphemes	Word vectors generated from nursing-care texts using word2vec	200	Class labels (C0: bad nursing-care, C3: very good)	Train: (2007-2008), Test: 2009
A6	Robert Sherwin et al. (2017)	NI	Vital signs, demographics, nursing assessments, specific laboratory values	NI	Sepsis status (sepsis, severe sepsis, septic shock, non-sepsis)	NI
A7	Shinichiroh Yokota et al. (2017)	Label encoding, Scaling	Sex, age, intensity of nursing care needs scores, admitting hospital ward, admitting department, month, day of the week	7	Fall occurrence	Holdout
A8	Jeungok Choi et al. (2018)	L1 regularization was used to prevent overfitting, Missing data were imputed	Physical health and illness, medication use, cognitive function, emotional health, sensory function, health behaviors, social connectedness, sexuality, and relationship quality variables	261	Depression (mild, severe)	10-fold cross-validation
A9	Stephen I. Gallant et al. (2018)	Text data vectorization, Text concatenation, Vector embeddings	Patient history, progress notes, lab reports, and other text notes from EHRs	12	Presence of severe sepsis	3-fold cross-validation
A10	Eliezer Bose et al. (2019)	NI	Problem-specific signs/symptoms, problem-specific ratings for knowledge, behavior, status, and service delivery variables including problem-specific intervention categories and targets	205	Maternal Risk Index (MRI) - dichotomous (high/low risk)	Train 50, Test 50
A11	Gerald C. Gannod et al. (2019)	Encoding	16 MDS 3.0 Section F preference items (8 personal care preferences, 8 activity preferences)	16	Resident preferences	Train 80, Test 20
A12	Steven G. Johnson et al. (2018)	TF-IDF, Handling of imbalanced datasets	Flowsheet row descriptions (60-120 characters of text)	800,000-1,000,000	Pain Information Model concepts	LOOCV

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A13	Zfania Tom Korach et al. (2019)	Normalization, Tokenization, Noise removal sequences	Weights of concepts extracted by topic modeling, vital signs, age, gender, note entry time	NI	Rapid response event risk	NI
A14	Jae Yung Kwon et al. (2019)	Dummy variable encoding, Standardization, Conversion to sparse matrices, Exclusion of missing data	Number of inpatient visits, discharge disposition, diagnosis codes, demographic information, medication types, lab test results, etc.	Over 50	30-day readmission	Train: 70, Test: 30
A15	Suzanne S. Sullivan et al. (2019)	Variable transformation, Feature selection, Categorization	Age, patient overall status (M1034), ADL Total Score, cancer diagnosis, frailty, oxygen use, medication management	36	12-month mortality (alive/not alive)	Train 70-90, Test 10-30
A16	Maxim Topaz et al. (2019)	Lowercasing, Punctuation removal, Stop word removal, Number removal	Words/phrases in the clinical notes	NI	Fall-related information categories	Train: 80, Test: 20
A17	Heather Brom et al. (2020)	NI	Age, sex, race, marital status, zip code, hospital length of stay, ED visits, 27 Elixhauser comorbidities, insurance type, admission source, discharge disposition	Over 35	30-day readmission	Train 50, Validation 25, Test 25
A18	Roschelle L. Fritz et al. (2020)	Segmenting sensor data around pain events, Labeling activities with ground truth, Filtering out Irrelevant sensor signals	Behavioral markers such as overall activity level, transitions, sleep quality and timing, grooming, and time spent out of home	550	Pain events (yes/no)	3-fold cross-validation
A19	Christopher M. Horvat et al. (2020)	Standardization, Missing data handled, Class imbalance, Feature selection	Vital signs, Laboratory data, Nursing assessments, Medication administration	NI	Clinical deterioration events	NI
A20	Renjie Hu et al. (2020)	Label encoding, Missing value handled	Demographics, nurse manager's leadership style, warmth and belonging climate, organizational trust, nurses' basic information, hospital error reporting system variables	68	report medication errors (ERREPQ1, ERREPQ2, ERREPQ3)	NI
A21	Mireia Ladios-Martin et al. (2020)	Missing value handled, Normalization, Dimensionality reduction	Medical service, days of oral antidiabetic/insulin therapy, ability to eat (Barthel scale), number of red blood cell units transfused, Hemoglobin range, PI present on admission, Illness severity (APACHE II score)	23	Pressure injury risk	Train: 59, Test: 41
A22	Soo-Kyoung Lee et al. (2020)	Normalization, Label encoding	Hours per resident day of staff, proportion of residents with psychiatric medications/urinary incontinence/aggressive behavior/cognitive decline, current number of residents, maximum capacity, turnover rates	57 (reduction 13)	Fall occurrence	5-fold cross-validation

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A23	Chen Liang et al. (2020)	Word stemming, Stop words removal, Alphabetic tokenizing, Lower cases transformation, TF-IDF	Text content of event reports	UHCS corpus: 2,064, WebM&M corpus: 2,037	Safety-related labels	Train: 90, Test: 10
A24	David S. Lindberg et al. (2020)	Imputation of missing values, Normalization, Selection of important features	Patient characteristics, admission information, assessment information, clinical data, staffing information	38	Fall occurrence	10-fold cross-validation
A25	Jung In Park et al. (2020)	Imputation using k-nearest neighbors algorithm, Binary transformation of categorical data, Cost-sensitive classification to address class imbalance	Age, gender, Charlson Comorbidity Index score, hospitalization within previous 6 months, immunosuppression, rationale for continued use of catheter, pre-existing catheter, lab result - glucose, length of hospital stay, total nursing hours per patient day, percent of direct care RNs with associate's degree, percent of direct care RNs with BSN, MSN, or PhD, percent of direct care RNs with specialty certification	13	Presence of HA-CAUTI	10-fold cross-validation
A26	Maxim Topaz et al. (2020)	Punctuation removal, Omitting stop words, Case normalization, Converting all words to lowercase, One-hot encoding of text data	Words and expressions from clinical notes	388	Hospitalization or ED visits	Train 70, Test 30
A27	Dana M. Womack et al. (2020)	Data normalization, Date mapping across systems, Feature selection using RFECV	Communication device minutes, medication delivery type percentages, Medication count skewness across RNs, Nurse call minutes, medications dispensed timing, patient continuity percentage, Mean medication count per patient	50 (reduction 8)	Unplanned overtime	8-fold cross-validation
A28	Ran An et al. (2021)	Standardization, PCA	Mechanical ventilation status, GCS score, levels of lactate, sodium, potassium, creatinine, blood urea nitrogen, mean arterial blood pressure, ventilatory support, cardiovascular support, renal support, hepatic support, neurologic support, nutritional support, specific interventions, and basic activities	16	Patient class (A, B, C based on severity and care needs)	7-fold cross-validation
A29	Linyan Chen et al. (2021)	Face detection, Facial landmarks location, Face alignment, Image Standardization, Normalization	Facial expression features extracted using histogram of oriented gradients	NI	Distress level (distressed/non-distressed)	Train 70, Test 30

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A30	Aaron Conway et al. (2021)	Normalization	Age, sex, ASA physical status classification, sleep apnea diagnosis, BMI, sedative/analgesic dose and type, total sedative dose, number of sedative doses, time since first sedation, time since previous sedative dose, previous respiratory state, duration of previous apneic event, time since previous apneic event, total number of apneic events	18	Prolonged apnea (>30 seconds)	10-fold cross-validation
A31	Alberto Garcés-Jiménez et al. (2021)	Data cleaning, Filtering, Individualization	Vital signs including body temperature, electrodermal activity, heart beat rate, oxygen saturation, blood pressure	11	Type of infectious disease (ARI, UTI, SSTI)	Train: 95, Test: 5
A32	Li Hannaford et al. (2021)	Handling of missing values, Standardization, Variable transformation	Demographics, high school GPA, residence type, term GPAs, cumulative GPAs, number of terms attended, nursing/non-nursing course GPAs	76	Graduation status (yes/no)	Train: 80, Test: 20
A33	Farinaz Havaei et al. (2021)	Handling of missing values, Standardization of factors	Psychological support, organizational culture, leadership expectations, civility & respect, psychological job fit, growth & development, recognition & reward, involvement & influence, workload management, engagement, balance, psychological protection, and physical safety.	13	Depression, anxiety, PTSD, Burnout (emotional exhaustion, depersonalization, personal accomplishment), Life satisfaction	Train: 70, Test: 30
A34	Elizabeth P. Howard et al. (2021)	Standardization	Demographics, functional status, cognitive status, mood/behavior, health conditions, diagnoses, medications/treatments	Over 100	Hospital readmission within 90 days	NI
A35	Mingyue Hu al. (2021)	Missing value handled	Age, instrumental activities of daily living, marital status, baseline MMSE score	4	Cognitive impairment (MMSE score < 18)	Train: 66.7, Validation: 16.7, Test: 16.7
A36	Oleksandr Ivanov et al. (2021)	Text normalization, Sentence tokenization, Word tokenization, Part-of-speech tagging, Chunking, Clinical term extraction	Age, vital signs, blood glucose, pain scores, GCS score, Morse Fall Scale score, sex, arrival mode, arrived from, family history, social history, risk factors, chief complaints, patient histories	26,332	Emergency Severity Index acuity score	5-fold cross-validation

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A37	Liuqi Jin et al. (2021)	Feature mapping, Data vectorization	Gender, Height, Weight, Degree, Ward, Age, Sensory perception, Moisture, Activity, Mobility, Nutrition, Friction and Shear, Body temperature, systolic blood pressure, diastolic blood pressure, Heart rate, Oxygen saturation, Blood sugar level, Hemoglobin, Albumin, Physical condition, surgery or trauma, duration of surgery, Incontinence, compulsive position, Braden PI score	26	Intervention type	10-fold cross-validation
A38	Jisu Kim et al. (2021)	NI	Age, sex, BMI, alcohol abuse, operation name, stroke history, GI bleeding history, MI history, coronary artery disease, hypertension, heart failure, diabetes, cancer, hepatic disease, atrial fibrillation, anemia, creatinine, labile INR, antibiotics, antiplatelet agents, herbs, NSAIDs, steroids.	22	Bleeding events (Yes/No)	NI
A39	Soo-Kyoung Lee et al. (2021)	Feature selection	Hours per resident day of director, proportion of bedridden residents, residents taking antidepressants or sleeping pills, residents with cognitive dysfunction, residents with urinary incontinence, residents with physical restraints, hours per resident day of certified nurse aide, number of current residents, proportion of Grade A facilities, retention rate of care workers	10	Occurrence of pressure injuries	NI
A40	Chia-Hui Liu et al. (2021)	Down-sampling	Demographics, admission diagnosis, self-care ability, physiological evaluation, surgical records, catheter records, focus charting, medication records, fall scale score	54	Fall occurrence	Train: 66.7, Test: 33.3
A41	Tamara G. R. Macieira et al. (2021)	Dimensionality reduction, Consolidation of classes, Feature selection	DA International domains, Nursing Outcomes Classification domains, Nursing Interventions Classification domains, keywords (family, self-care assistance, monitoring)	29	Palliative care categories (family, well-being, mental comfort, physical comfort, mental, safety, functional, physiological)	Train:66.7, Test:33.3

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A42	Takuro Nagata et al. (2021)	Image cropping, Color calibration using CasMatch color patch, Segmentation into smaller image regions (SLIC superpixel segmentation)	RGB channels, HSV channels, Lab* color space, erythema index, melanin index, shape features of wound regions.	11	STAR classification category (levels 1, 2, or 3 for wound severity).	LOOCV
A43	Gojiro Nakagami et al. (2021)	Handling missing value	Age, sex, ward type, eating function, incontinence factors, communication functions, physical function restrictions, skin conditions, respiratory and cardiac symptoms, difficulty in activities of daily living	Over 50	Occurrence of pressure injuries	Train:70, Test:30
A44	Wenyu Song et al. (2021)	Handling of missing values (excluded patients with >25% missing data), Feature engineering	Demographics, GCS score, level of consciousness, gait/transferring, activity, pain score, diabetes, peripheral vascular disease, spinal cord injury, stroke, anemia, albumin, blood urea nitrogen, chloride, potassium, sodium, creatinine, hemoglobin, white blood cell count, platelet blood count	28	Occurrence of pressure injuries	Train 80, Test 20
A45	Rumei Yang et al. (2021)	Mean imputation for missing data, Min-max normalization	Age, gender, mental and physical health status, healthcare access, smoking, drinking, exercise habits, chronic conditions	68	Occurrence of Fall (yes/no)	Train 75, Test 25
A46	Huaqiong Zhou et al. (2021)	Imputation of missing values	Age, sex, admission type, length of stay, insurance status, referral source, socioeconomic status, distance from hospital, ICU stay, general anesthesia, previous hospital usage, significant lab results, social history, language, completeness of discharge documentation, delay in discharge summary issuance.	40	30-day readmission	10-fold cross-validation
A47	Yanhong Dong et al. (2022)	Data transformation, Over-sampling	Depression, anxiety, stress, intrusion, avoidance, hyperarousal scores	6	Whether a healthcare worker is a nurse (binary: 1 = nurse, 0 = other healthcare workers)	Train 80, Test 20

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A48	Farinaz Havaei et al. (2022)	Dummy coding	Psychological support, organizational culture, leadership expectations, civility and respect, psychological job fit, growth and development, recognition and reward, involvement and influence, workload management, engagement, balance, psychological protection, physical safety	13	Quality of nursing care, Patient safety grade, Workplace recommendation	Train 70, Test 30
A49	Tingting Hu et al. (2022)	Missing value handling, Feature selection	Age, height, weight, body mass index, gestational age, number of cesarean sections, number of abortions, Bishop score, fetal weight, amniotic fluid index, amniotic fluid contamination, fetal head circumference, fetal abdominal circumference, fetal biparietal diameter, fetal bone length, maternal uterine height, maternal abdominal circumference, fetal membrane status, labor analgesia	18	Outcome of induction of labor (Success/Failure)	NI
A50	Shuai Jin et al. (2022)	Zero-centering, Scaling, Dummy encoding, Median imputation	Age, Charlson Comorbidity Index score(CCI), chemotherapy, port-Cath, NSAID, bed, VTE history, WBC, plaster, and D-dimer.	10	Cancer-associated DVT occurrence	Train 70, Test 30
A51	Mireia Ladios-Martin et al. (2022)	Missing data handled, SMOTE	Fall prevention, age, days of psychometrics treatment, sex, incontinence, chronic obstructive pulmonary disease, family support, diabetes, gait, and various drug treatments	Model A: 13, Model B: 22	Fall occurrence	Train 50, Test 50
A52	Young Ji Lee et al. (2022)	Bag-of-words representation	Words from posts (text features)	NI	Needs categories (physical, psychological/ emotional, social, health information)	10-fold cross-validation
A53	Anup Kumar Mishra et al. (2022)	Standardization, Handling missing data	ADL, IADL, MMSE, GDS, SF-12 Physical and Mental Component Scores (PCS, MCS), Functional Ambulation Profile (FAP), Gait Speed, Fall History	9	Fall occurrence	5-fold cross-validation

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A54	Kyoung Ja Moon et al. (2022)	NI	Age, sex, birthdate, diagnosis, disease severity, number of comorbidities, pain, pain medicine use, abnormal blood urea nitrogen, dehydration, water-electrolyte imbalance, nutritional imbalance, hypoxia, infection, sleep disorder, surgery with general anesthesia, edema, age ≥ 65 years, visual impairment, hearing impairment, cognitive impairment, level of consciousness, changes in consciousness, severity of delirium	24	Delirium risk (high, moderate, low)	10-fold cross-validation
A55	Nikhil Padhye et al. (2022)	Linear detrending, removal of circadian rhythm	Entropy measures, scaling exponent, Braden scale score, and demographic/vital signs	4	Occurrence of pressure injuries	LOOCV
A56	Dongni Qian et al. (2022)	NI	Gender, age, BMI, marital status, insulin level, blood glucose, pediatrician function, skin thickness, blood diabetic pressure, pulse	10	Recurrence of diabetes	Train 70, Test 30
A57	Javier Rojo et al. (2022)	Feature selection	31 items from the ENCS form	14 (first phase), 25 (second phase)	Functional profile of aging adults (five continuous values between 0 and 100)	NI
A58	Jiyoun Song et al. (2022)	Missing value handled, SMOTE	Socio-demographic factors, care related factors, medical conditions, risk for hospitalization, sensory status, integumentary status, ADLs/IADLs, and risk factors from clinical notes	75	Hospitalization or ED visits	Train 90, Test 10
A59	Tobias R. Spiller et al. (2022)	Down-sampling	Sex, age, 56 items from ePA-AC nursing assessment	58	Delirium presence (based on Delirium Observation Scale score ≥3)	Train: 71.9, Test 28.1
A60	Katie Walker et al. (2022)	Data Standardization, One-hot encoding, Handling Standard deviations, missing values	Triage category, arrival time, arrival method, ambulance data, and previous patient wait times	19	ED wait time (triage-to-provider).	Train: (2017-2018)Test: (2019)
A61	Melyana Nurul Widyawati et al. (2022)	NI	Age, parity, height, inter-pregnancy interval, hemoglobin levels, upper arm circumference, prior diseases, and bleeding history	8	Pregnancy risk level (no risk, low risk, moderate risk, high risk)	10-fold cross-validation
A62	Jie Xu et al. (2022)	Missing value handled	Demographics, admission cause classification, clinical laboratory data, medical history, Braden score components, GCS	31	Occurrence of pressure injuries	Train 70, Test 30

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A63	Olga Yakusheva et al. (2022)	Missing value handled, Feature selection	Nurse HPPD for RN non-overtime, RN overtime, non-RN staff; patient demographics, unit characteristics, hospital characteristics	141 (reduction 60)	30-day readmission	Train 70, Test 30
A64	Ayla İ rem Aydın et al. (2023)	Video Standardization	11 Facial Action Units (AU4, AU6, AU7, AU9, AU10, AU12, AU17, AU20, AU25, AU26, AU45)	11	Pain intensity score (0-10 scale)	Train 67, Test 33
A65	Rui CHEN et al. (2023)	NI	Underlying lung disease, smoking history, serum albumin ≤ 35 g/L, radiotherapy history	4	CIP occurrence.	NI
A66	Ya-Huei Chen et al. (2023)	Excluding unreasonable data, SMOTE	Age, gender, body measurements, vital signs, chronic diseases, special diseases, surgeries, drug types, pain score, fall history	46	Fall occurrence	5-fold cross-validation
A67	Pei-Yu Dai et al. (2023)	Feature selection	Gender, age, ICU days, APACHE II score, respirator mode, respiratory parameters, creatinine, lactate, blood sugar, blood pressure, pulse, respiration rate, catheter type, medication records	NI	RASS-based classification	10-fold cross-validation
A68	Odai Y. Dweekat et al. (2023)	Single imputation, Normalization, Dummy variables	Braden scale components, demographics, medical history, diagnosis, labs, medications, medical devices	98	Occurrence of pressure injuries	Train:80, Test:20
A69	Juliet Edgcomb et al. (2023)	NI	Safety screening questions, nursing safety interventions, diagnoses, medications, labs, prior care use	NI	Presence or absence of suicidality	10-fold cross-validation
A70	Ajeet Gajra et al. (2023)	NI	Age, sex, race/ethnicity, primary tumor type/stage, socioeconomic factors, clinical information	NI	Risk score for ED visit or admission within the next 30 days	NI
A71	Farinaz Havaei et al. (2023)	NI	Psychological support, organizational culture, leadership expectations, civility and respect, psychological job fit, growth and development, recognition and reward, involvement and influence, workload management, engagement, balance, psychological protection, physical safety	13	Type II violence (patients/visitors), Type III violence (organizational employees)	Train 70, Test 30
A72	Sharon Hewner et al. (2023)	Recoding missing values, Binary recoding of conditions, Creation of dummy variables	Age, gender, race, primary language, hospital visit count, medical conditions, social indicators.	25	Patient clusters/ segments	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A73	Sunho Im et al. (2023)	Imputation of missing values, Feature selection with >50% missing data	Respiratory pattern, pulse, RASS, pupil reflex, consciousness level, device count, urine output, respiratory sound, SpO2, respiratory rate, ventilator use, blood pressure, body temperature, etc.	14-78	ICU mortality (yes/no)	Train 80, Test 20
A74	Junglyun Kim et al. (2023)	Missing value handled	Gender, education, economic status, marital status, diabetes mellitus, hypertension, arthritis, physical disability, age, social disconnectedness, depression, loneliness, pain, sleep disturbance, self-burden, life meaning	17	Suicidal ideation (yes/no)	Train 70, Test 30
A75	Seong-Kwang Kim et al. (2023)	Label encoding, Standardization, Normalization, Dummy coding, Missing value handled, Standard deviation removal	Age, sex, team, grade, income, dormitory use, marital status, distance from home to workplace	8	Nurse turnover (resignation status)	Train 80, Test 20
A76	Hyungbok Lee et al. (2023)	Dichotomization	Triage level, sex, age, visit day, visit time, visit type, referring area, referring hospital type, Inter-hospital communication, severe emergency disease, emergency operation or angiography, ICU admission, ward admission, re-transfer, number of consultations, number of diagnoses, neoplasms diseases, circulatory diseases, respiratory diseases	19	Length of stay in the emergency department	10-fold cross-validation
A77	Hyungbok Lee et al. (2023)	One-hot encoding	ED visit factors: Age, visit day, visit time, visit route, visit mode, visit type, severity level, mental status, comorbidities, chief complaint	14	Workplace violence occurrence	10-fold cross-validation
			ED stay factors: Expression of dissatisfaction, consultation of other specialties, average daily number of patients, average daily length of stay			
A78	Lin-Lin Lee et al. (2023)	Dataset structuring, Deletion of non-related wound information	Spectral data of the wound area	16	Occurrence of pressure injuries	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A79	Lingjuan Li et al. (2023)	Feature selection	Age, sex, headache, visual disturbance, polyuria, weakness, tumor diameter, tumor type, calcification, Puget classification, Surgical approach, stalk preservation, resection extent, preoperative serum sodium, cortisol, thyroxine, free thyroxine	17	Postoperative severe hypernatremia	Train: 186, Test: 48
A80	Pei-Hung Liao et al. (2023)	NI	Age, gender, body mass index, hypertension, glaucoma, liver disease, sciatica, diabetes mellitus, hyperlipidemia, asthma, rheumatoid arthritis, gastroesophageal reflux disease, cardiovascular disease, cardiovascular medications, antihypertensives, liver disease medications, stomach medications, hypoglycemic agents, hormone medications, hypolipidemic agents, hypnotics, analgesics, traditional Chinese medications	23	Sarcopenia risk	NI
A81	Sarah R. Martha et al. (2023)	Missing value handling, Standardization	Male, age, race, comorbidities, stroke risk, BMI, lipid panel on admission, Medications (home), NHISS on admission, NHISS on discharge, tPA administration, stroke location, TICI score, LKW to reperfusion time	13	Acute ischemic stroke status	LOOCV
A82	Aruna Jothi Shanmugam et al. (2023)	NI	Demographic variables, knowledge about self-medication, reason for self-medication, type of self-medication, duration of self-medication, sources of obtaining self-medication	6	Frequency of self-medication practice	NI
A83	Araceli Rodríguez Vico et al. (2023)	Feature selection, Resampling (SMOTE), Feature selection	Gender, age, cardinal presentation, Previous diseases, previous medications, NIHSS score, MRS score, Imaging diagnosis, fibrinolysis, thrombectomy	11	Neurological status at discharge	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A84	Zeping Yan et al. (2023)	Dummy variables, Feature selection	Age, sex, BMI, education level, residence, employment status, marital status, household income, type of medical insurance, smoking status, hypertension, diabetes, heart disease, duration of knee pain before surgery, disease diagnosis, ASA grade, surgical site, anesthesia, operative duration, intraoperative blood loss, NRS at rest, activity, PCS, TSK, GSES, HSS	26	Chronic post-surgical pain (Yes/No)	Train 70, Test 30
A85	Metin Yildiz et al. (2023)	Normalization	Age, gender, marital status, education, monthly income, years of working in nursing, care providing status, ethnocentrism, and intercultural sensitivity	9	Health tourism awareness level	Train 70, Test 30
A86	Ying Zhou et al. (2023)	Normalization, Synthetic Minority Over-sampling Technique (SMOTE), Tomek Links for class balancing	Speech features, Facial features, Text features	Over 90	Emotional states (normal, depression only, anxiety only, apathy only,	10-fold cross-validation
A87	Maryam Zolnoori et al. (2023)	Lowercasing, Removing punctuation, Symbols, Numbers, Stop words removal (for TF-IDF)	TF-IDF, LIWC, Word2Vec, UMLS concepts, N-grams, POS tagging	Multiple features (TF-IDF: 6338, Word2Vec: 200, LIWC: 93, UMLS: 673)	Speaker type (patient vs nurse)	5-fold cross-validation
A88	Young-Taek Park et al. (2024)	Normalization, Encoding of categorical variables, Standard deviation removal using Cook's distance	Years of operation, number of doctors, number of nurses, number of beds, percentage of physician specialists, percentage of nurses among nursing staff, number of CTs, number of MRIs, local population, local households, hospital location, foundation type	12	Total length of inpatient stay (TLOS) and total number of outpatient visits (TNOV)	Train 80, Test 20
A89	Cynthia ABI KHALIL et al. (2024)	Standard deviations handling, Missing value handling	Demographic variables, comorbidities, nutritional status, laboratory results	34	HAPI risk level (High/Low)	Train 80, Test 20
A90	Yaser ALQARRAIN et al. (2024)	Standardization, Factorization of variables, Multiple imputation, One-hot encoding, Normalization, Dimensionality reduction	Nursing assessment variables, demographic factors	267	Healthcare-Acquired Urinary Tract Infection (HAUTI) occurrence (yes/no).	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A91	Komal Aryal et al. (2024)	Data cleaning, Removal missing data, Variable factorization	Age, sex, comorbidities, polypharmacy, nutrition risk, daily decision making, ADL, CPS, pain	20	30-day COVID-19 mortality	10-fold cross-validation
A92	Ranjana Chavan et al. (2024)	Standardization, Feature selection	Academic performance, clinical competencies, extracurricular activities, personal preferences, CGPA, logical thinking, numeric ability, English proficiency, coding ability, standing arrears.	NI	Placement outcome (successful/ unsuccessful)	Train 70, Validation 15, Test 15
A93	Xiaomei Chen et al. (2024)	One-hot encoding, Min-max normalization	Operation time, intraoperative corticosteroids administration, preoperative skin protection measures, preoperative skin conditions, gender, an hepatic phase time, disease category, cholinesterase, albumin, BMI	10	Occurrence of pressure injuries	Train 70, Test 30
A94	Colum Crowe et al. (2024)	Standardization, Feature computation, Down-sampling	Cadence, steps per day, minutes active, percent time inactive, percent time in (low/medium/ high) activity, Stride velocity peak, cadence average/median, stride length, max (60/20/5/1), peak performance index, gait speed, Lyapunov exponents, sample entropy, hurst exponent, correlation Dimension, detrended fluctuation analysis	26	Pre/post intervention classification	Train 80, Test 20
A95	Tian Dai et al. (2024)	Multiple imputation, LASSO regression, Ordinal ranking, One-hot encoding	BMI, operation duration, history and status of COPD, prealbumin, TNM staging, stoma site, TRAM, CRP, ASA classification, stoma diameter, and others	48	Parastomal hernia occurrence (Yes/ No)	Train 70, Test 30
A96	Martha Duarte et al. (2024)	Normalization, Variable scoring	Drugs alcohol interpretation, drugs alcohol points, drugs alcohol did you have a dr, sex, aid, tobacco use are you a, sexual history have you had an, drugs alcohol intake, visit type, sexual history are you experience, age, race, tobacco use are you exposed to, drugs alcohol have you used dr, social living with someone, sexual history have you had no, sexual history have you had se, language, ethnicity, social are you exposed to haze, social highest level of education, social are you involved in any, marital status, insurance name	24	PHQ-2 score (depression screening result)	Train 80, Test 20

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A97	Yu-Fang Guo et al. (2024)	Dummy variables, Random number sequence for data split	Job crafting, leisure crafting, demographics	16	Burnout	Train 70, Test 30
A98	Rui Jin et al. (2024)	One-hot encoding	Gender, language spoken, marital status, additional contact listed, age, diagnosis, body part under assessment, type of assessment, wound being assessed	9	Likelihood of receiving antimicrobial treatment.	NI
A99	Arisa Kawashima et al. (2024)	Under sampling, Feature selection	Pain NRS, impact thermometer, fall history, HbA1c, distress thermometer, age, prothrombin time, symptom NRS, total protein, means of transport to hospital, anxiety, emetogenic risk of chemotherapy regimen, phosphorus, sex, total cholesterol, gamma-glutamyl transferase	16 (reduction 5)	Specialist palliative care (SPC) needs	Train 80, Test 20
A100	Yeonju Kim et al. (2024)	Under-sampling	Age, gender, type of hospital admission, type of ICU, length of stay in ICU, frequency of vital signs measurement, frequency of nursing note documentation, Body temperature, monitoring-related nursing notes, heart rate	10	ICU mortality	Train 80, Test 20
A101	Ju Hee Lee et al. (2024)	Missing value handling, Feature selection	Age, gender, medical department, length of ICU stays, urinary disturbance, sensory perception impairment, smoking history, BMI, RASS, KPCS, albumin, hemoglobin, lactic acid, use of medical devices	182	Occurrence of pressure injuries	Development cohort (January 2018 to December 2020), Validation cohort (January 2021 to May 2022).
A102	Pin-Chieh Lee et al. (2024)	Data quality assessment, Deletion of incomplete data, Standardization	Urine zinc, urine cadmium, urine cadmium/creatinine, aflatoxin, free radicals, blood cadmium, PAH, diethyl phthalate (uMep), di(2-ethylhexyl) phthalate (uMEHP)	9	Lung cancer occurrence	Train 70, Test 30
A103	Renee C. B. Manworren et al. (2024)	Frame extraction, Labeling	NFCS facial action features	8	Pain events (yes/no)	Train 70, Test 30

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A104	Ninon Girardon da Rosa et al. (2024)	Data selection, Transformation, Handling missing data, k-fold cross-validation	Nursing care, nursing diagnoses, vital signs records, diet prescriptions, medication prescriptions, solution prescriptions, hemotherapy prescriptions, oxygen therapy prescriptions, nursing-collected exam requests, specialist consultation requests, educational practices records, age, sex, education, hospitalization date	15	Nursing workload categories (minimum, intermediate, semi-intensive, intensive)	5-fold cross-validation
A105	Jihye Kim Scroggins et al. (2024)	Lowercase conversion, Punctuation/Symbol/Number/Stop word removal, Stemming, Lemmatization	TF-IDF, bigram, POS tagging features	NI	Health problems identified in verbal communication	Train 70, Test 30
A106	Lu Shao et al. (2024)	Down-sampling, LASSO feature selection	Balance, grip strength, fatigue, fall history, age, comorbidity.	6	Fall occurrence	NI
A107	Madeleine Stanik et al. (2024)	Missing value handled (k-nearest neighbors), Data balancing (SMOTE, ENN, SMOTEENN)	Demographics, treatments (physical therapy, occupational therapy, speech therapy, recreational therapy, psychological therapy, medications), physical condition (daily activities, mobility, balance), behavior (mood, pain, delirium)	120	Seizure occurrence within 14 days after infection	Train 80, Test 20
A108	Imam Tahyudin et al. (2024)	Missing value handling (zero imputation), One-hot encoding for categorical variables, Label encoding	Length of hospitalization, age, gender, history of cardiovascular disease, post medical history, past medical history, previous stroke history, HB (Hemoglobin), HT (Hematocrit), LEU (Leukocyte), TR (Thrombocyte), NLR (Neutrophil-Lymphocyte Ratio), Cholesterol Total, HDL, TG (Triglyceride), LDL, stroke location	17	Mortality (survival vs. non-survival)	Train 80, Test 20

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A109	Metin Yıldız et al. (2024)	NI	Gender, marital status, having immigrant acquaintances, foreign language knowledge level, experience abroad, having migrant patients to care for, age, participation in intercultural interaction sub-dimension, respect for cultural differences sub-dimension, self-confidence in intercultural interaction sub-dimension, the sub-dimension of enjoying intercultural interaction, attention to intercultural interaction sub-dimension	14	Xenophobia level	Train 70, Test 30
A110	Cheng Yu et al. (2024)	Multiple imputations for missing data, Standardization	Job stress dimensions (nursing profession and work issues, time allocation and workload, working environment and equipment problems, patient care issues, management and interpersonal problems), childhood adversity, sociodemographic and work-related variables	Over 20	Sleep quality (PSQI score)	Train 70, Test 30
A111	Wei Zhang et al. (2024)	Multiple imputation for missing values, Standardization, Feature selection	Demographic data, health-related risk factors, and chronic disease risk factors	46	Frailty status (binary: frailty vs. non-frailty)	Train 75, Test 25
A112	Maryam Zolnoori et al. (2024)	Lowercasing, removing extra punctuation, speaker diarylation, Missing values handling	Linguistic features, patient demographics, health risk factors from EHRs, clinical notes, and audio features	Over 100	ED visits and hospitalizations within 60-day period post-admission	LOOCV
A113	Yunping Zhang et al. (2022)	Missing value handled	Facility characteristics, staffing variables, resident characteristics	36	Readmission rate and length of stay (LOS)	NI
A114	Goran Erfani et al. (2024)	K-means clustering, Standardization	VAX scale factors (mistrust of vaccine benefit, worries about unforeseen future effects, concerns about commercial profiteering, preference for natural immunity), sociodemographic factors, work-related factors, health-related factors, media usage data	25	Vaccination hesitancy levels	NI
A115	Tae Youn Kim et al. (2006)	EM imputation, frequency-based replacement, attribute reduction	Braden subscales: activity, friction/shear, sensory perception; clinical factors: edema, indwelling catheter, nutrition consult, wheelchair use, extra nursing care	8	Pressure ulcer occurrence	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A116	In Sook Cho et al. (2011)	Missing value handling, variable selection based on missing rates	Demographics, clinical observations, physiological variables	36	Pressure ulcer occurrence	Train: 67, Test: 33
A117	Yoko Setoguchi et al. (2016)	Data cleansing to exclude missing data	Sex, age, disease, BMI, items of NNS-A, NNS-B	NI	Pressure ulcer occurrence	10-fold cross validation
A118	Mikyung Moon et al. (2017)	Variable transformation, binarization, table joining, filtering	Demographic, disease/treatment, healthcare provider characteristics	830	Pressure ulcer occurrence	10-fold cross-validation
A119	Pacharmon Kaewprag et al. (2017)	Feature selection through statistical analysis, data cleaning	Braden scale measurements, medications (72 categories), diagnoses (ICD-9 codes)	86	Pressure ulcer occurrence	Train: 67, Test: 33
A120	Xiaohong Deng et al. (2017)	NI	Age, ICU length of stay, temperature, blood pressure (systolic, diastolic, mean), oxygen saturation, hemoglobin, albumin, arterial blood gas analysis, mechanical ventilation use, diabetes, infection, Braden Scale scores, smoking status, edema, fecal incontinence	20	Pressure ulcer occurrence	10-fold cross-validation
A121	Hong-Lin Chen et al. (2018)	Random division of dataset	Age, gender, disease category, weight, smoke, diabetes, perioperative albumin levels, surgery duration, CPB duration, postoperative MV, Vasoactive agents Intraoperatively, Vasoactive agents postoperatively, Corticosteroids perioperative	NI	Pressure ulcer occurrence	Train: 70, Test: 30
A122	Hsiu-Lan Li et al. (2019)	NI	Age, sex, history of pressure injuries, disease type (cancer/non-cancer), length of hospitalization, mental status, excretion, activity/mobility, local skin sensation, skin condition/circulation, nutrition	11	Pressure ulcer occurrence	NI
A123	Seul Ki Park (2020)	NI	Braden Scale items, Scott Triggers tool items, laboratory results, type of anesthesia, comorbid conditions	22	Pressure ulcer occurrence	NI
A124	Sookyung Hyun et al. (2021)	Missing data handling, data cleaning	Age, gender, weight, diabetes, vasopressor, isolation, endotracheal tube, ventilator episode, Braden score, ventilator days	10	Pressure ulcer occurrence	NI

(Continued on the next page)

Supplementary Table 4. Continued

Index	Author (Year)	Data Preprocessing Techniques	Predictor Variables	Number of Predictor Variables	Target Variable	Data Split Ratio
A125	Ji-Yu CAI et al. (2021)	NI	Age, gender, disease category, weight, duration of surgery, duration of cardiopulmonary bypass procedure, perioperative corticosteroid administration, use of intraoperative vasoactive agents, use of postoperative vasoactive agents	9	Pressure ulcer occurrence	NI

ADI: Area Deprivation Index; ADL: Activities of Daily Living; APACHE II: Acute Physiology and Chronic Health Evaluation II; ARI: Acute Respiratory Infection; ASA: American Society of Anesthesiologists; AU: Facial Action Units; CGPA: Cumulative Grade Point Average; CIP: Critical Illness Polyneuropathy; COPD: Chronic Obstructive Pulmonary Disease; CPB: Cardiopulmonary Bypass; CPS: Cognitive performance Scale; CRP: C-Reactive Protein; CT: Computed Tomography; DVT: Deep Vein Thrombosis; ED: Emergency Department; ENCS: Elderly Core Nursing Set; ENN: Edited Nearest Neighbours; ePA-AC: Electronic Patient Assessment-Acute Care; ESI: Emergency Severity Index; FAP: Functional Ambulation Profile; GCS: Glasgow Coma Scale; GDS: Geriatric Depression Scale; GI: Gastrointestinal; GSES: General Self-Efficacy Scale; HA-CAUTI: Healthcare-Associated Catheter UTI; HAPI: Hospital-Acquired Pressure Injury; HbA1c: Hemoglobin A1c; HOG: Histogram of Oriented Gradients; HPPD: Hours Per Patient Day; HSS: Hospital for Special Surgery; HSV: Hue, Saturation, Value color space; IADL: Instrumental Activities of Daily Living; ICD-9: International Classification of Diseases, Ninth Revision; ICU: Intensive Care Unit; INR: International Normalized Ratio; KPCS: Korean Patient Classification System; Lab*: CIELAB color space; LASSO: Least Absolute Shrinkage and Selection Operator; LKW: Last Known Well; LOOCV: Leave-One-Out Cross-Validation; LOS: Length of Stay; MCS: Mental Component Summary; MI: Myocardial Infarction; MMSE: Mini-Mental State Examination; MRI: Magnetic Resonance Imaging; MRI: Maternal Risk Index; MRS: Modified Rankin Scale; MV: Mechanical Ventilation; NFCS: Neonatal Facial Coding System; NI: No Information; NIHSS: National Institutes of Health Stroke Scale; NNS-A / NNS-B: Nursing Needs Score A / Nursing Needs Score B; NRS: Numeric Rating Scale; NSAIDs: Nonsteroidal Anti-Inflammatory Drugs; PAH: Polycyclic Aromatic Hydrocarbons; PCA: Principal Component Analysis; PCS: Physical Component Summary; PHQ-2: Patient Health Questionnaire-2; PI: Pressure Injury; PSQI: Pittsburgh Sleep Quality Index; PTSD: Post-Traumatic Stress Disorder; RASS: Richmond Agitation-Sedation Scale; RFECV: Recursive Feature Elimination with Cross-Validation; RN: Registered Nurse; SF-12: Short Form 12 Health Survey; SLIC: Simple Linear Iterative Clustering; SMOTE: Synthetic Minority Oversampling Technique; SMOTEENN: SMOTE + Edited Nearest Neighbors; SpO2: Oxygen Saturation; SPC: Specialist Palliative Care; SSTI: Skin and Soft Tissue Infection; TF-IDF: Term Frequency-Inverse Document Frequency; TICI: Thrombolysis in Cerebral Infarction; TLOS: Total Length of Stay; TNM: Tumor, Nodes, Metastases staging; TNOV: Total Number of Outpatient Visits; tPA: tissue Plasminogen Activator; TRAM: Transverse Rectus Abdominis Musculocutaneous; TSK: Tampa Scale for Kinesio phobia; UTI: Urinary Tract Infection; VAX: Vaccine Attitudes Index; VTE: Venous Thromboembolism; WBC: White Blood Cell.