An Ensemble Dynamic Model and Bio-Inspired Feature Selection Method-based Decision Support System for Predicting Multiple Organ Dysfunction Syndrome in the ICU

Anas Maach[1], El Houssine El Mazoudi[2], Jamila Elalami[3], Noureddine Elalami[4]

LASTIMI, High School of Technology in Sale, Mohammed V University , Rabat, Morocco[1,3]

CISIEV. Faculty of Science and Technology, Cadi Ayyad University, Marrakech, Morocco[2]

LASTIMI. Mohammedia School of Engineers, Mohammed V University, Rabat, Morocco[4]

Abstract—Multiple Organ Dysfunction Syndrome (MODS) is one of the most common and severe conditions affecting patients admitted to intensive care units (ICUs). It is characterized by the simultaneous failure or dysfunction of at least two organ systems. Although no specific remedy for MODS has been identified to date, early diagnosis and adequate organ support can significantly improve patient outcomes. Identifying patients at risk of developing MODS in the ICU is challenging. Currently, several methods are used for this purpose, including scoring systems like SOFA and MOD Score, as well as machine learning-based approaches. However, these methods often have limitations. Some require invasive features, making them complex to use in a smart healthcare system. Others suffer from a lack of performance due to various problems, which can potentially lead to unreliable predictions. Feature selection can improve ML models' performance. Recently, bio-inspired feature selection techniques have shown promise in improving the performance of machine learning methods in many domains, but their effectiveness in MODS prediction has not yet been evaluated. Additionally, research on early MODS prediction, particularly utilizing time-series data and dynamic ensemble methods, remains limited. To fill this gap, the present research used state-of-the-art machine learning algorithms, namely dynamic ensemble techniques, to predict patients at risk of developing MODS in the ICU. Dynamic ensembles are new methods that select an ensemble of the best-performing models for every new test case. We compared the performance of these models with full features and with feature selection. Three nature-inspired meta-heuristic optimization models, namely the binary bat algorithm (BBA), grey wolf optimization (GWO), and genetic algorithm (GA), were evaluated to select the optimal feature subset. The models were built using non-invasive patient features and time-series data from the first 12 hours of ICU admission. The results showed that feature selection significantly improved the performance of dynamic ensemble models. Notably, the METADES model, employing grey wolf optimization for feature selection, demonstrated the best performance in terms of accuracy(96.5%), F1 score (96.4%), precision (97.2%), recall (95.7%), and area under the ROC curve (AUC) (98.4%). These findings highlight the potential and effectiveness of our approach for early MODS prediction in ICUs.

Keywords—*Ensemble dynamic model; MODS prediction; decision support system; Bio-Inspired feature selection*

I. INTRODUCTION

Multiple Organ Dysfunction Syndrome (MODS) is widely recognized as a primary cause of death in critically ill patients, affecting 11% to 40% of adults admitted to intensive care units (ICUs) [1]. Accordingly, the high mortality rate of this syndrome, ranging from 44% to 76%, underlines its seriousness. MODS typically arises in response to severe illness or injury, often as a result of conditions like sepsis, severe trauma, major surgery, or prolonged shock. It involves the simultaneous dysfunction or failure of at least two organ systems, such as the heart, lungs, liver, and kidneys, etc. While the dysfunction is typically acute and severe, there is potential for reversibility, especially with prompt identification and treatment of underlying causes or triggers [2].

Despite extensive research, effective treatments for MODS remain elusive. Current interventions have not adequately controlled the excessive immune response or facilitated organ recovery. This has led to invasive organ support as the primary treatment approach in ICUs [1]. Additionally, a survey by the American Hospital Association (AHA) revealed that there are over 6,300 intensive care units in 3,200 acute care hospitals in the United States, providing a total of 94,000 ICU beds [3]. Consequently, the shortage of medical staff in ICUs exacerbates work pressures, affecting patient care quality and potentially leading to oversight of crucial changes in patient conditions [4]. Therefore, rapid diagnosis becomes essential for optimal resource allocation to the neediest patients. It's important to note that the implementation of early-phase management strategies, including a resuscitation approach focused on damage control and scoring systems, has contributed to an increased survival rate among injured patients upon admission to intensive care. Hence, fundamental aspects of MODS treatment involve early identification and support of organ functions. Several scoring systems have been developed to assess the severity of Multiple Organ Dysfunction Syndrome (MODS) and predict outcomes using clinical parameters. Among these, the SOFA score (Sequential Organ Failure Assessment) [5] is commonly utilized. The SOFA score is designed to monitor and predict the progression of organ failure by assessing the function of six organ systems: cardiovascular, liver, respiratory, coagulation, renal, and neurological [6]. Furthermore, each

organ system is assigned a score ranging from 0 to 4, as shown in the Table I. A higher score reflects more severe failure. Organ failure is typically identified by a SOFA score exceeding 2 in one of the six assessed organ systems [7]. The total SOFA score is the sum of these individual scores, ranging from 0 to 24. However, the SOFA score is a complex tool that necessitates meticulous patient evaluation and the continuous collection of numerous parameters. Consequently, it may exhibit variability in predicting the outcome of MODS.

TABLE I. SOFA SCORE PARAMETERS

Organ Systems	SOFA	Score
Respiratory	PaO2/FiO2 and ventilation	0 to 4
Coagulation	Platelet number	0 to 4
Hepatic	Bilirubin	0 to 4
Cardiovascular	Blood pressure and vasopressor use	0 to 4
Central nervous system	Glasgow Coma Scale	0 to 4
Renal	Creatinine urine output	0 to 4
Aggregate	Calculated daily	0 to 24

In recent years, there has been a gradual increase in research on intelligent intensive care units, with a focus on monitoring and risk prediction. By leveraging modern scientific and technological advancements such as 5G communication technologies, the Internet of Things, and big data analysis, coupled with machine learning techniques[4], these innovations hold promise in predicting the likelihood of an individual developing MODS in the ICU.

Many researchers have employed ensemble classifiers for clinical classification problems, with promising results [8] [9] [10][11]. Additionally, they have also explored Dynamic Ensemble Selection (DES) models [12][13] [14] that select an ensemble of classifiers dynamically for each test data item. This allows DES to identify patterns in complex domains like biomedicine, credit scoring, and handwriting recognition instead of relying on a single classifier for the entire dataset. While existing research on predicting Multi-Organ Dysfunction Syndrome (MODS) in ICU patients suggests room for improvement, we investigated DES techniques for this purpose. Our study compares them with diverse feature selection methods utilizing nature-based algorithms. Feature selection helps alleviate the "curse of dimensionality," facilitating faster, simpler, and potentially more performant machine learning models. Analyzing relevant studies [15], [16], [17], [18], [19], we identified the Binary Bat Algorithm (BBA), Genetic Algorithm (GA), and Grey Wolf Optimization (GWO) as effective and prevalent feature selection algorithms. Our study employed these bio-inspired feature selection methods.

Following feature selection, we employed seven state-of-the-art DES models (META-DES, DESP, KNORA-U, DESKNN, KNORA-E, MCB, and KNOP) for classification. We comprehensively evaluated their performance using metrics like F1-measure, recall, sensitivity, precision, accuracy, and ROC curve analysis to select the optimal classifier. Additionally, Area Under the Receiver Operating Characteristic (AUROC) curve analysis assessed the impact of feature selection.

Our main objective is to develop a decision support system capable of accurately classifying and predicting patients at risk of having MODS in the ICU using only non-invasive features and time-series data from the first 12 hours after ICU admission. This system has the potential for integration into a smart healthcare monitoring system for intensive care units[8], as illustrated in Fig. 1.

The remainder of the document is structured as follows: Section II provides an overview of the current state of research in MODS, followed by Section III, which outlines the proposed methodology. Section IV presents the experimental results and subsequent discussion. Section V addresses the limitations of this study and suggests avenues for future research, while Section VI serves as the conclusion.

II. RELATED WORKS

To date, extensive research efforts have been dedicated to investigating the diagnosis of Multiple Organ Dysfunction Syndrome (MODS).Various medical and artificial intelligence-based methods have been employed, yielding significant results. This section undertakes a thorough examination of the literature concerning MODS diagnosis and related triggers, such as sepsis:

Bowen et al. [20] proposed an approach based on machine learning for predicting multi organ dysfunction syndrome (MODS) recovery in pediatric patients with sepsis. The authors highlight the lack of effective predictive models for early recovery from MODS in this patient group. The study introduces a novel methodology that anticipates the transition from MODS to milder states, utilizing datasets from Swiss and U.S. pediatric sepsis cohorts. The model demonstrated promising performance, achieving approximately 79.1% AUROC and 73.6% AUPRC during internal validation and 76.4% AUROC and 72.4% AUPRC during external validation. The suggested approach exhibits the potential for integration into electronic health record systems, thereby aiding in patient evaluation and prioritization within pediatric sepsis care. Furthermore, the researchers employ SHAP values to elucidate pivotal recovery factors as identified by the model. The study also explores associations between predicted outcomes and factors such as pathogens, infection sites, and age groups, contributing to an enhanced interpretation of the model's predictions.

Guanjun et al. [21] proposed a study to develop models to predict multiple organ dysfunction syndrome (MODS) among trauma patients utilizing noninvasive predictors alone. Traditional methods of predicting MODS are invasive and difficult to implement in a pre-hospital environment. The study uses records from 2319 patients and employs seven machine learning methods to create real-time MODS prediction models. A comparison was made between the models and the four conventional scoring systems. The best-performing model is based on LightGBM (LGBM) and Adaboost, achieving a high AUC of 0.959 when using all parameters. Even when reducing the parameters to non-invasive ones only, the LGBM model still outperformed traditional scoring systems, with an AUC of 0.940. The study concludes that the accurate, real-time prediction approach using non-invasive features is superior to conventional scoring systems, which could facilitate early diagnosis and improve patient survival rates in the pre-hospital setting.

Chang et al. [22] proposed an advanced approach to the challenge of predicting and preventing multiple organ dysfunction syndrome (MODS), The study used machine learning
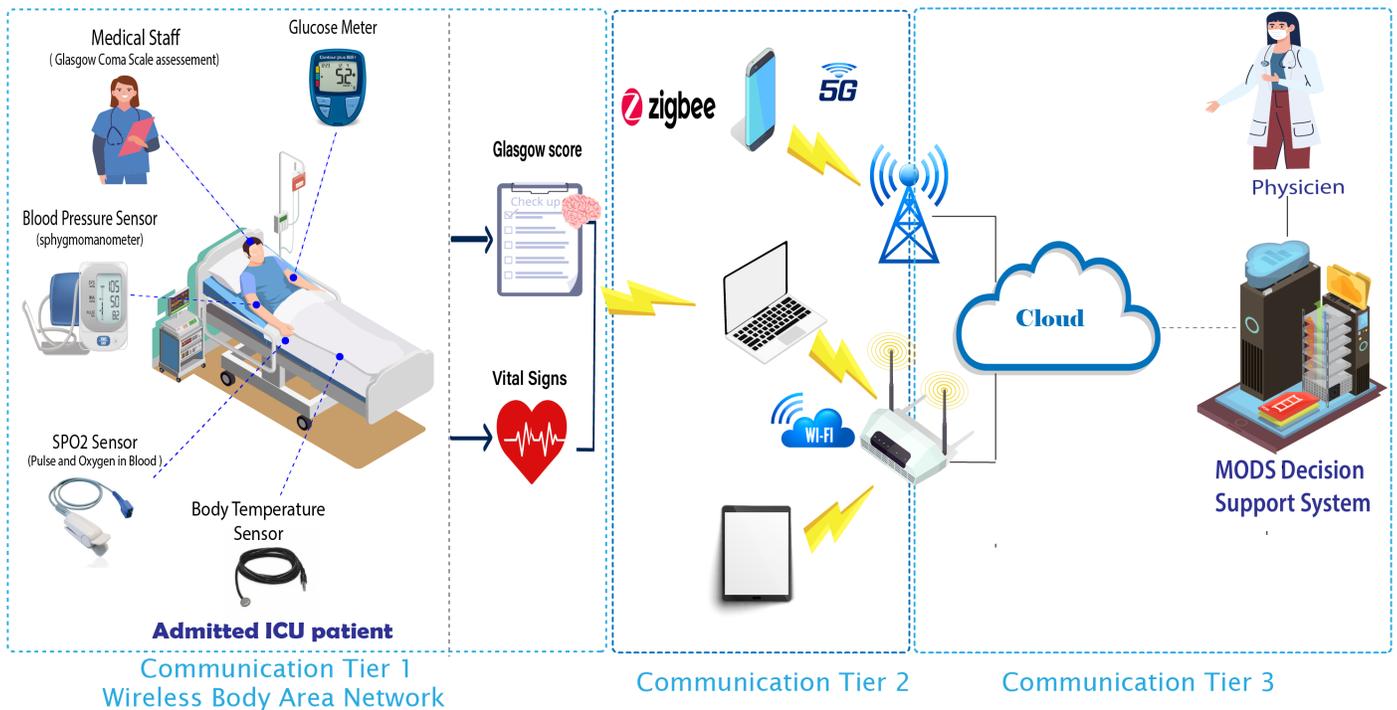
Fig. 1. General architecture of the predictive system for MODS in ICU.

algorithms to create early warning models for MODS risk assessment. The researchers developed a customizable model called SuperLearner by integrating several machine learning approaches, which resulted in high robustness in detection and favorable assessment measures such as sensitivity and accuracy. In addition, the deep neural network (DWNN) model showed excellent performance, with a notable AUC of 0.960 on the MIMIC-IV test set. This result highlighted the exceptional predictive power of the DWNN model in the context of MODS prediction. The research introduced additive Kernel-SHAP and various DiCE (counterfactual explanations) to interpret the results of the predictions and recommend interventions, enabling the prediction of MODS risk 12 hours in advance. The study utilized clinical features, scoring characteristics, and database data for model training, showcasing the potential application value of their approach for early MODS prediction and intervention. The integration of Q-learning for model selection and the combination of SuperLearner and SubSuperLearner structures were innovative contributions. The researchers also introduced a utility score for comprehensive performance assessment and incorporated DiCE to facilitate automatic intervention recommendations for improved practicality.

Tunc et al. [23] introduced a solution for monitoring sepsis-related symptoms and the condition of organ systems without the need for lab tests. Their work proposed the Deep SOFA-Sepsis Prediction Algorithm (DSPA), which combines features from Convolutional Neural Networks (CNNs) with the Random Forest (RF) algorithm to predict Sequential Organ Failure Assessment (SOFA) scores of patients diagnosed with sepsis using only seven vital signs collected in the Intensive Care Unit (ICU). They evaluated their model using the MIMIC III dataset and achieved a mean absolute error (MAE) of

0.65, a correlation coefficient (CC) of 0.86, and a root-mean-square error (RMSE) of 1.23 in predicting SOFA scores at the onset of sepsis. Their model demonstrated superior performance compared to traditional machine learning and deep learning models in regression analysis. Furthermore, they showcased strong classification performance, achieving an area under the curve (AUC) of 0.982 for predicting early sepsis, surpassing previous studies. The proposed framework offers a non-invasive and timely approach for predicting sepsis and monitoring organ states.

Alexis et al. [24] conducted a study on multiple dysfunction syndrome in children after congenital heart surgery involving cardiopulmonary bypass (CPB). The study involved 306 surgical patients under the age of 18 and collected biomarkers and clinical information. The model, called PERSEVERE-CPB, incorporated the level of interleukin 8 (IL-8) 12 hours after bypass surgery, the change in serum chemokine ligand 3 (CCL3) between 4 and 12 hours, and the infant's age category. PERSEVERE-CPB was able to efficiently stratify patients into categories of low, intermediate, and high risk for the development of persistent MODS, demonstrating the potential for targeted interventions and improved outcomes through the identification of high-risk patients. The discriminative performance of the model was comparable to reference tools such as the STAT model and the PRISM III score, with an AUROC of 0.86 (95% CI 0.81; 0.91) for discrimination between patients with and without persistent MODS. After 10-fold cross-validation, the PERSEVERE-CPB model maintained good performance, with a corrected AUROC of 0.75 (95% CI 0.68-0.84).

This concise overview underscores the significant interest within the scientific community regarding the diagnosis

and prediction of Multiple Organ Dysfunction Syndromes (MODS). Despite the limited number of studies utilizing machine learning models for MODS prediction and the scarcity of research involving non-invasive features, resulting in the complexity of implementing these models into an intelligent decision support system for early MODS prediction, it is clear that machine learning tools have yet to achieve widespread application in MODS diagnostic systems. This holds especially true in developing countries, where the mortality rate due to MODS remains alarmingly high, leaving considerable room for improvement. In this study, we introduce a novel approach based on a dynamic ensemble model and an advanced feature selection method for selecting the optimal feature set. Setting it apart from other methods, this approach is straightforward to implement, utilizes non-invasive features to forecast the risk of MODS occurrence in the ICU, and has exhibited outstanding performance in the detection and prediction of Multiple Organ Dysfunction Syndrome (MODS).

III. Proposed Methodology

In this study, our goal is to develop a Decision Support System based on a predictive model for predicting which patients are at risk of developing MODS in the ICU. To achieve this, we compared seven state-of-the-art Dynamic Ensemble Selection models (DES) that assess the skill of individual classifiers from a classifier pool. The most skilled classifier, or a set containing the most skilled classifiers, is then used to predict the correct label for a given test sample. We formulated a classification problem aimed at predicting the risk of a patient developing MODS based on extracted data gathered in the initial 12 hours following their admission to the intensive care unit.

The evaluation of the base classifiers was carried out through the cross-validation technique, and the DES models were assessed using a validation test set to estimate the ability of our model to generalize outside its trained dataset. Throughout our research, we explored various models and architectures, testing different feature sets by applying three nature-based optimization techniques.

The proposed methodology is outlined and visually represented in Fig. 2.

A. Study Design and Datasource

In this study, we employed MIMIC-III, a medical database containing anonymized records from more than 46,520 patients admitted to the intensive care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts. The data spans from 2001 to 2012. The database has received ethical approval for use and is managed by the Massachusetts Institute of Technology's Computational Physiology Laboratory (MIT) under the PhysioNet-accredited Health Data 1.5.0 license. This extensive database comprises 26 tables containing a diverse range of data, such as demographic information, vital sign measurements, laboratory test results, medical procedures, medication records, caregiver notes, imaging reports, and mortality data upon discharge. These data are interconnected using key identifiers such as subject-ID, hadm-ID, and ICUSTAY-ID. In order to ensure patient confidentiality, a rigorous de-identification process was applied, aligning with the Health

Insurance Portability and Accountability Act (HIPAA) standards in the United States. This process involved the removal of personally identifiable information, such as patient names, phone numbers, and specific dates. Additionally, a date-shifting method was employed to preserve temporal intervals in the data. We obtained approval to extract data from this database under (Record ID: 53063368).

B. Data Pre-Processing

These steps are designed to improve the overall quality of the selected dataset. The MIMIC III dataset contains a number of issues, such as outliers, missing values, etc. This can be the consequence of a sensor or data transfer failure, an error in data storage, etc. Building a model with such poor, incomplete data is regarded as the major factor behind underperforming models. The initial process of data preprocessing involves converting raw data to a convenient and useful format. This stage involves three distinct steps: cleaning the data, data transformation, and data reduction. Data cleaning focuses on resolving problems associated with missing data and anomalies. The data transformation phase aims to reshape the data so that it is more adaptable for data mining. Commonly used transformation techniques include attribute selection, normalization, etc. Finally, data reduction avoids the complications associated with processing large datasets. The next subsections deal with the measures taken to deal with these data problems.

1) Data extraction: Data were extracted from the MIMIC-III dataset (v1.4). Apache Spark software was used to extract baseline features (subject ID, ID of ICU stay, age, gender), vital signs, and non-invasive features as shown in Table II from patients meeting the criteria using SQL (Structured Query Language) as shown in the cohort selection Fig. 3, and to extract pertinent features to compute the SOFA scoring system.

Our decision to employ Apache Spark for data extraction from MIMIC-III was driven by its remarkable capability to efficiently handle large volumes of distributed healthcare data. Spark excels in distributed processing and provides a unified suite of tools, rendering it highly suitable for extracting and processing complex medical data at scale. Its fault tolerance and capacity to distribute workloads across a cluster of machines ensure the reliability and performance necessary for analyzing clinical data from MIMIC-III.

In addition to its data extraction capabilities, Apache Spark also offers a versatile environment for further data analysis, enabling researchers and healthcare professionals to gain valuable insights from this extensive medical dataset. The combination of MIMIC-III and Apache Spark has proven to be a powerful solution for in-depth healthcare analytics and research.

2) Missing data handling: A thorough understanding of data is of crucial importance when analyzing data in the healthcare context. The challenges inherent in this field call for proper management of missing data. Although the simplistic method of removing missing values is commonly used, it has the notable disadvantage of leading to a loss of significant information, thus reducing the number of data instances available for model training. In response to this problem, various strategies have been proposed for filling in missing values using alternative records, such as forward filling and
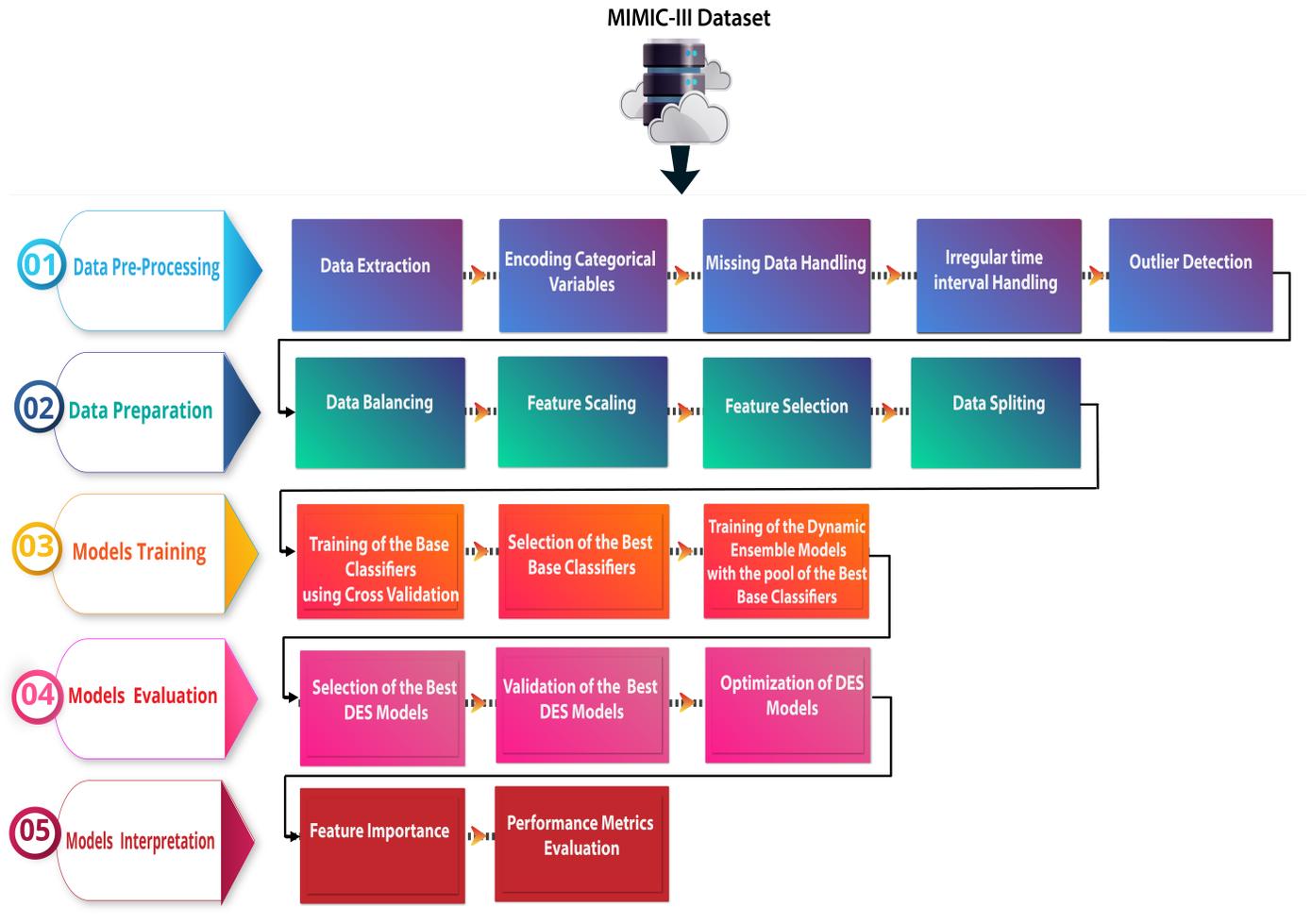
Fig. 2. The proposed approach.

the use of K-Nearest Neighbors. During the MIMIC III data pre-processing phase, a substantial proportion of crucial data (between 40 and 55%) is unfortunately lost. However, given its importance to the forecasting process, deleting this data was not an option. Faced with this challenge, we decided to select cases with two or more values in each measure, and then apply the forward filling method to impute the remaining missing values.

3) Outliers detection: In our study, special emphasis was placed on handling outliers within the medical dataset. The detection of outliers was performed using the interquartile range method to avoid the indiscriminate removal of records. Outliers were addressed in two steps: first, they were replaced with null values, considered missing, and then imputed using the forward-filling method.

4) Irregular time interval: In the MIMIC III dataset, the recording of vital signs occurred at irregular intervals, varying from measurements taken every few minutes to every few seconds. This irregularity in time intervals presented a challenge for machine learning techniques that typically operate with uniformly sampled data. To overcome this challenge, we aggregated patient vital sign observations, consolidating them into a single record every hour. This aggregation involved

incorporating key statistical measures, including the standard deviation, mean, minimum value, maximum value, and count of all measurements within each hourly interval. As a result, each record now contains consistent values. In addressing further irregularities within the temporal intervals of time-series data, particularly with regard to balancing measurements for patients diagnosed with MODS in the dataset, we implemented a targeted approach. The initial step involved organizing the data by MODS patient ID and timestamp. Subsequently, each patient underwent individual processing to tackle irregular measurements by either eliminating excess or filling gaps with randomly generated dates. The handling of null values was achieved through the forward and backward filling methods, strategically replacing missing values based on predefined criteria. These meticulous steps ensure the consistency of measurements across temporal datasets, ultimately enhancing the quality and reliability of future analyses.

C. Data Preparation

1) Data balancing: The imbalance of classes is one of the most well-known and crucial issues that can influence the performance of machine learning algorithms. This issue occurs when classes are unequally represented. In unbalanced data, majority classes dominate minority classes. Consequently,

TABLE II. FEATURES USED IN THIS STUDY

Feature	Definition
Age	Admission age of the patient
Gender	Gender of the patient
Height	The measurement of a patient's vertical size. Used for assessing body proportions.
Weight	The measurement of a patient's mass. Used for various health assessments, including medication dosages.
Diastolic blood pressure	The pressure in the arteries when the heart is at rest. It is an essential indicator of cardiovascular health.
Systolic blood pressure	The pressure in the arteries when the heart contracts. Important for assessing cardiovascular health and blood flow.
Fraction inspired oxygen	The proportion of oxygen in the air or a gas mixture that is being inhaled. Important for assessing respiratory function and oxygen delivery.
Glucose	The level of glucose in the blood. A critical indicator of glycemic control and metabolic health.
Heart Rate	The number of heartbeats per minute. Crucial for assessing cardiac function and rhythm.
Oxygen saturation	The percentage of hemoglobin in the blood that is saturated with oxygen. It is important for evaluating respiratory function and oxygenation.
Respiratory rate	The number of breaths taken per minute. Essential for monitoring respiratory health and efficiency.
Temperature	The measurement of a patient's body heat. Crucial for monitoring body temperature and detecting fever or hypothermia.
pH	The measure of the acidity or alkalinity of the blood. Essential for evaluating acid-base balance and overall metabolic health.
Mean blood pressure	The average pressure refers to the average pressure in the arteries throughout one cardiac cycle. It serves as a crucial indicator of overall blood pressure.
Glasgow Coma Scale Eye Opening	used to evaluate a patient's level of consciousness by assessing their eye response.
Glasgow Coma Scale Motor Response	used to evaluate a patient's level of consciousness based on their motor response.
Glasgow Coma Scale Verbal Response	used to evaluate a patient's level of consciousness by assessing their verbal response.

Fig. 3. Cohort selection diagram from MIMIC dataset.
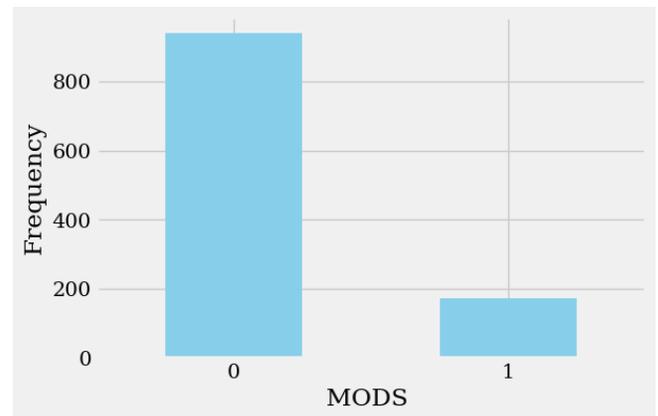
Fig. 4. Distribution of classes before applying the Smote technique.

since there are not enough instances of the minority class, an imbalanced classification has the disadvantage that a model cannot effectively learn the decision boundary, and machine learning approaches have a higher probability of classifying each new observation in the majority class. Consequently, the issue of unbalanced data can lead to the misclassification of minority classes. However, there is a significant need for an effective method that could address the class imbalance problem. In this study, the minority class has 172 samples, while the majority class has 940 samples, resulting in a ratio of 5.4:1, as depicted in Fig. 4. Thus, we employed an unsupervised technique, namely the Synthetic Minority Oversampling Technique (SMOTE)[25], to address the class imbalance issue in the datasets as depicted in Fig. 5.

2) Feature scaling: Feature scaling is a preprocessing technique used in statistics and machine learning to normalize the values of different features in a dataset. Often, datasets contain features that vary widely in terms of size, units, and range. The goal is to adjust the scales of features so that they are comparable, and no single feature dominates others due to its units or magnitude. The range of intensive care unit data points considered in this study is very diverse, and therefore it is necessary to perform feature scaling to minimize any effects on model performance. In our study, we have chosen Z-score scaling as our standardization method.

This decision stems from the need to make our features robust to outliers, a crucial aspect in the context of our data. Unlike other methods, such as normalization, Z-score scaling minimizes the impact of extreme observations, ensuring a more balanced scaling of features. Moreover, this approach facilitates the interpretation of results, especially in the context of linear models, by providing directly comparable coefficients. By prioritizing standardization, our goal is to optimize the stability and convergence of machine learning models, thereby contributing to more reliable analyses and robust results within our study.
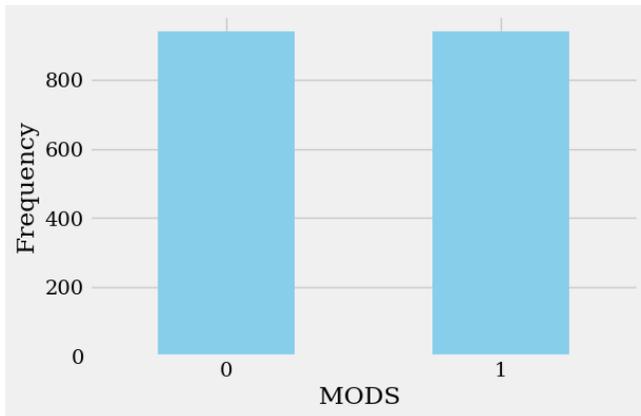
Fig. 5. Distribution of classes after applying the Smote technique.

3) Feature selection: Feature selection is an important component of feature engineering and plays a key role in improving the capability of machine learning algorithms [26]. The primary contribution of the proposed approach lies in its capability to carefully select a subset of features of interest from the set of extracted features, resulting in significantly improved prediction results. This diagnostic model employs an optimal feature selection approach. The primary objective is to emphasize relevant features while reducing the number of features to address redundancy issues. Overall, this methodology aims to minimize the feature set during the construction of the predictive model, leading to a reduction in computational costs and an enhancement in overall model performance. Recent studies highlight the effectiveness of Nature-Inspired optimization feature selection approaches, contributing to a notable increase in model performance and efficiency. The feature selection algorithms employed in this study involve:

a) Grey Wolf Optimization (GWO): The Grey Wolf Optimizer (GWO) [27] is a nature-inspired optimization algorithm. This algorithm simulates the cooperative and hierarchical hunting strategy of wolves, and its key steps are as follows:

Surround the Prey (Initialization): The algorithm begins by randomly placing a population of wolves in the search space, each representing a potential solution.

Hunting Behavior (Fitness Evaluation): Each wolf's fitness is evaluated using a fitness function, measuring its alignment with optimization goals and reflecting its hunting success in finding the optimal solution.

Hierarchy: Alpha, Beta, and Delta (Leadership Selection): Wolves are sorted based on their fitness levels. The top three wolves are identified as alpha, beta, and delta, establishing a leadership hierarchy within the pack.

Update Positions (Pack Movement): The positions of wolves are adjusted using a formula inspired by the social behavior observed in wolf packs. Alpha, beta, and delta play pivotal roles in directing the movement of other wolves toward potentially optimal solutions.

Exploration and Exploitation (Hunting Strategy): The hierarchy ensures a balance between exploration and exploitation. Alpha, beta, and delta lead the exploration, while other wolves follow, exploring around their positions to discover potential solutions.

Surrounding the Prey (Optimization): The algorithm iterates through these stages, progressively refining the positions of wolves. This mimics the way a wolf pack surrounds prey during a hunt, improving the chances of finding the optimal solution.

Criteria for Completion (End of Hunt): The algorithm continues these stages for a defined number of iterations or until a predefined termination criterion is met. The final positions of the wolves represent the optimized solutions.

b) Binary Bat Algorithm (BBA): The Binary Bat Algorithm (BBA), detailed in [28], is an optimization algorithm inspired by the echolocation behavior of bats. It is specifically engineered for tackling binary or combinatorial optimization problems. It simulates bats' use of ultrasonic pulses for navigation and prey location, incorporating features like frequency and intensity modulation, as well as global and local search mechanisms. BBA has demonstrated versatility and effectiveness in addressing various optimization problems since its inception.

The Binary Bat Algorithm (BBA) comprises the following key steps:

Sonar Scanning (Initialization): Initialize a population of binary bats randomly within the search space, representing potential feature subsets.

Fitness Echo (Objective Function Evaluation): Evaluate the fitness of each bat solution using a task-specific objective function for feature selection.

Leadership Hierarchy (Alpha, Beta, and Delta Bats): Establish a leadership hierarchy by designating the top-performing bats as alpha, beta, and delta.

Echo-Driven Movement (Flight Adjustment): Adjust the positions of bats, guided by alpha, beta, and delta, influencing the exploration of potential optimal feature subsets.

Adaptive Echolocation (Exploration and Exploitation): Maintain a balanced exploration and exploitation strategy, with alpha, beta, and delta leading exploration and other bats following suit.

Echo-locative Refinement (Optimization Iterations): Iteratively refine bat positions, mimicking the echolocation process and progressively enhancing the chances of identifying an optimal feature subset.

Termination by Convergence (End of Echolocation): Continue iterations until a predefined convergence criterion is met or a specified number of iterations is completed. The final bat positions represent the optimized feature subsets.

c) Genetic Algorithm (GA): The Genetic Algorithm (GA) [29] introduced the idea of using a population-based search inspired by biological evolution to solve optimization problems. The concept has since evolved, and various adaptations of genetic algorithms have been proposed and applied to different domains, including feature selection in machine learning.

The Genetic Algorithm (GA) comprises the following key steps:

Initialization: Generate an initial population of potential solutions, each representing a binary feature subset.

Evaluation: Assess the fitness of each solution based on a fitness function, evaluating its performance with the selected features.

Selection: Choose individuals from the population to act as parents for the next generation, favoring those with higher fitness.

Crossover (Recombination): Combine genetic material from selected parents to create new offspring.

Mutation: Incorporate minor random alterations to select individuals to uphold genetic diversity.

Replacement: Substitute a portion of the current population with the newly generated offspring.

Termination Criteria: Verify if a termination criterion has been satisfied, which could entail reaching a maximum number of generations or attaining a designated fitness threshold.

Result Extraction: Extract the final chromosome or feature subset from the population as the optimized set of features.

D. Machine Learning Algorithms

1) Dynamic Ensemble Selection Models (DES): are a promising and relevant technique belonging to the category of MCS approaches. Using base classifiers, they dynamically choose the most skilled classifiers for every new test item being classified, with each classifier being competent in a local 'feature space' region. These approaches have shown superior results compared to traditional ensemble methods that combine the results of base classifiers.

a) META-DES (Dynamic Ensemble Selection using Meta-Learning): [30] is a machine learning algorithm designed for dynamic ensemble selection in the field of classification. Its main objective is to approach classification dynamically by treating it as a meta-problem involving determining whether a particular classifier, chosen from a set of classifiers, is competent enough to accurately classify specified test data. This process involves two main steps. First, meta-features such as a posteriori probability for every label, the classifier's overall local accuracy, a vector indicating the difficulty of classifying neighboring instances, and the classifier's confidence based on the perpendicular distance separating the input sample from its decision boundary are derived. Subsequently, meta-classifiers exploit these meta-features to predict the ability of the selected classifier to provide accurate predictions for the designated test data. The classifiers identified by the meta-classifiers are then merged to construct a set of classifiers for the specified test data. META-DES essentially adopts a meta-perspective on classification, striving to dynamically choose the best-performing classifiers for a particular task, based on extracted meta-features.

b) DESP (Dynamic Ensemble Selection with Probability): [31] is an algorithm designed for dynamically selecting the best classifiers from an ensemble by eliminating those deemed incompetent. This is done by evaluating the performance of a single classifier against a random one. The performance given by the random classifier is determined by taking 1/M, with M being defined as the total number of classes that exist in the dataset. Classifiers are dynamically selected for each test data set on the basis of their performance relative to the performance achieved by the random neighborhood classifier selected for the test data set. If the performance of a classifier outperforms the random one, it is deemed suitable for selection into the ensemble for this particular test data set. If no classifier is selected, all classifiers in the ensemble are chosen for the given test dataset. In summary, the algorithm aims to create a dynamic ensemble of classifiers by eliminating incompetent ones and favoring those with better performance than a random classifier in a specified neighborhood.

c) KNORA-U (K-Nearest-Neighbor Algorithm for Dynamic Classifier Selection): The KNORAU algorithm, as outlined in [32], is designed to enhance the accuracy of classifying test samples. It utilizes the concept of k-nearest neighbors (KNN) by identifying the K closest neighbors for each test sample based on distances in feature space. KNORAU then selects classifiers from the initial pool that have accurately classified at least one neighbor among the K nearest, thereby forming a sample-specific ensemble. The prediction of the test sample's label employs the majority vote rule within this ensemble, with vote weights determined by each classifier's past performance in the K-nearest neighborhood. In essence, KNORAU strategically leverages classifier performance within the vicinity of the K-nearest neighbors to improve classification accuracy.

d) Dynamic Ensemble Selection KNN (DESKNN): The DES-KNN method, as introduced in [33], is an ensemble classifier selection algorithm aimed at identifying an optimal set from an initial group of classifiers. It employs diversity and accuracy as selection criteria. Initially, the algorithm identifies the most accurate classifiers within the competence region of a given test dataset. It then proceeds to select the most diverse classifiers among the most accurate ones using a measure known as the double-fault measure. Percentage-based selection, informed by prior research, dictates the proportion of classifiers chosen based on their accuracy and diversity. These percentages have been determined based on the superior performance observed in previous studies.

e) KNORA-E (K-Nearest-Neighbor Algorithm for Ensemble): [32] is a dynamic ensemble selection approach. It aims to choose a set of classifiers from a pool that can accurately classify all K nearest neighbors in a test dataset within a specific training set. The selection process is dynamic, eliminating classifiers that fail to classify at least one nearest neighbor correctly. Once the classifier set is identified, it is used for majority voting in subsequent classifications, following the majority voting rule. If an ensemble isn't found with the initial K value, KNORAE progressively adjusts the K value downward until at least one classifier set is identified. In summary, KNORAE, based on DES, seeks to select a robust set of classifiers capable of correctly classifying the nearest neighbors of a test point within a specific training set.

f) Multiple-Classifier Behavior (MCB): The MCB method[34] involves determining the competence region of a new test sample using the behavioral knowledge space (BKS) and the accuracy of the local classifier. Output profiles are

generated for the test sample and its competence region. The similarity between the output profiles of the test sample and those of its skill region is measured. Samples with similarities below a specified threshold are ignored, allowing the size of the proficiency region to be adjusted. The skill of the base classifier is assessed on the basis of its classification accuracy in this adjusted skill region. If a selected classifier has a significant performance advantage over the others (with a difference in skill exceeding a predetermined threshold), it is used for classification. Alternatively, all classifiers are then combined using the majority vote rule.

g) k-Nearest Output Profiles (KNOP): The KNOP method[35] consists of the selection of classifiers that have classified one or more samples within the expertise area of the sample being queried. The region of competence is determined by analyzing the decisions made by the base classifier, known as output profiles. Rather than considering the feature space, the degree of similarity that exists between the queried sample and the validation sample is evaluated through the decision space. Every classifier chosen is allocated a number of votes equivalent to the actual number of samples in the skill region where it accurately predicts the label. The cumulative votes of all core classifiers are then combined to produce the final ensemble decision.

IV. RESULTS

A. Performance Evaluation Metrics

In our evaluation of the proposed approach, we gauge its performance using essential performance metrics. Accuracy assesses the overall correctness of a classification model by comparing correctly predicted instances to the total. Precision quantifies the relevance of positive predictions, while recall evaluates the model's capability to identify actual positives. AUC (Area Under the ROC Curve) serves as a binary classification metric, representing the area under the curve that plots the true positive rate against the false positive rate at various thresholds.

a) Accuracy: Accuracy stands as a prevalent evaluation metric utilized to assess the overall performance of a classification model. It denotes the ratio of correctly predicted instances to the total number of instances within the dataset.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Instances}}$$

b) Precision: Precision serves as a metric that quantifies the proportion of true positive predictions out of all positive predictions made by the model. It offers a measure of how many of the predicted positive instances are indeed relevant.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

c) Recall (Sensitivity or True Positive Rate): The recall, also referred to as sensitivity or true-positive metric, evaluates the ratio of true-positive instances that are correctly classified by the model.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

d) AUC (Area Under the Receiver Operating Characteristic Curve): AUC, commonly used for binary classification problems, is a performance metric that represents the area under the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate at various threshold levels.

The AUC is calculated by integrating the ROC curve:

$$\text{AUC} = \int \text{TPR(FPR)} \, d\text{FPR}$$

e) F1-score (F1-measure): The F1-score is a measure of a model's accuracy, balancing both precision and recall.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

B. Machine Learning Models Analysis

In this section, we delve into the analysis of the performance of various dynamic ensemble selection methods for the classification and prediction of Multiple Organ System Dysfunction (MODS) using vital signs from the initial 12 hours in the Intensive Care Unit (ICU). Experiments were conducted on a computer with an Nvidia GeForce MX 350 graphics card, an Intel Core i7-10700T processor, and 16 GB of RAM based on scikit-learn 1.1.2 in Python 3.10.8.

We conducted four experiments, each in search of the combination of the best classifier and the most efficient feature selection method for MODS prediction, respectively. Our approach involved testing various state-of-the-art dynamic ensemble models with and without bio-inspired feature selection methods. Model selection was based mainly on comparing their performance statistically. As shown in Fig. 7, four distinct results were reported for the tested models: without feature selection as shown in Fig. 7d, with the genetic algorithm as shown in Fig. 7a, with the binary bat algorithm as shown in Fig. 6b, and with the grey wolf optimization as shown in Fig. 7c.

The dataset was split into two subsets: 70% for training and 30% for testing. The training set was utilized to perform optimization and training of baseline classifiers using cross-validation, while the test set was used to evaluate the performance of dynamic ensemble selection (DES) models based on various metrics. Seven state-of-the-art DES models were applied: META-DES, DESP, KNORA-U, DESKNN, KNORA-E, MCB and KNOP. In addition, three bio-inspired feature selection algorithms were used to identify the most appropriate feature subset: GWO, BBA and GA. The models were applied to both the entire feature set and the selected features.

a) Analysis of Results using All Features: In this section, we investigate the performance of dynamic ensemble models with full features. Table III provides details of the results achieved by the ML models on several evaluation measures. We summarise the results as follows: Using DESKNN and KNOP with full feature sets produced minor

TABLE III. PERFORMANCE METRICS FOR ENSEMBLE MODELS WITH DIFFERENT FEATURE SELECTION TECHNIQUES

Ensemble Models	Feature Selection	Accuracy	Precision	Recall	F1	AUC
KNORA-E	GWO	**0.945**	**0.933**	**0.957**	**0.945**	**0.973**
	BBA	0.853	0.951	0.741	0.833	0.864
	GA	0.906	0.942	0.863	0.901	0.915
	Without FS	0.926	0.943	0.905	0.924	0.965
METADES	GWO	**0.965**	**0.972**	**0.957**	**0.964**	**0.984**
	BBA	0.915	0.934	0.891	0.912	0.959
	GA	0.89	0.927	0.843	0.883	0.935
	Without FS	0.936	0.944	0.927	0.936	0.973
KNORA-U	GWO	**0.942**	**0.929**	**0.957**	**0.943**	**0.971**
	BBA	0.84	0.916	0.744	0.821	0.867
	GA	0.902	0.929	0.868	0.898	0.917
	Without FS	0.915	0.922	0.905	0.913	0.964
DESKNN	GWO	**0.94**	**0.924**	**0.957**	**0.94**	**0.968**
	BBA	0.836	0.907	0.744	0.817	0.871
	GA	0.898	0.922	0.868	0.894	0.909
	Without FS	0.891	0.869	0.92	0.894	0.951
MCB	GWO	**0.94**	**0.93**	**0.949**	**0.94**	**0.967**
	BBA	0.837	0.91	0.744	0.818	0.863
	GA	0.898	0.922	0.868	0.894	0.896
	Without FS	0.909	0.906	0.912	0.909	0.955
DESP	GWO	**0.934**	**0.927**	**0.939**	**0.933**	**0.978**
	BBA	0.838	0.913	0.744	0.82	0.852
	GA	0.898	0.922	0.868	0.894	0.909
	Without FS	0.915	0.932	0.894	0.912	0.962
KNOP	GWO	**0.941**	**0.928**	**0.954**	**0.941**	**0.974**
	BBA	0.835	0.897	0.751	0.818	0.869
	GA	0.873	0.875	0.868	0.871	0.917
	Without FS	0.886	0.86	0.92	0.889	0.951

performance (accuracy=0.891, precision=0.869, recall=0.92, F1-score=0.894, and AUC=0.951) and (accuracy=0.886, precision=0.86, recall=0.92, F1-score=0.889, and AUC=0.951), respectively. MCB , DESP and KNORA-U improved their performance by approximately 3% compared with KNOP. KNORA-E improved its performance by approximately 1.1% compared with KNORA-U. The highest performance was achieved with METADES (accuracy = 0.936, precision = 0.944, recall = 0.927, F1 score = 0.936 and AUC = 0.973). Fig. 6d depicts the AUC and ROC curves of the models with full features. The METADES model achieves the highest AUC (0.973), while the KNOP model achieves the lowest AUC (0.951). Fig. 7d shows the radar plot for models with full features, and places the METADES model in the outperforming category.

b) Results Analysis using the Grey Wolf Optimization (GWO) for Feature Selection: In this section, we investigate the performance of the Ensemble Dynamic Models with selected features by the GWO. Table III provides details of the results achieved by the ML models on several evaluation measures. We summarise the results as follows: Using DESP with selected feature sets produced minor performance (accuracy = 0.934, precision = 0.927, recall = 0.939, F1-score = 0.933 and AUC = 0.978) . MCB and DESKNN improved performance with about 0.6% above DESP. KNORA-U and KNORA-E improved their performance by approximately 0.2-0.3% above MCB and DESKNN. The highest performance was achieved with METADES (accuracy=0.965, precision=0.972, recall=0.957, F1-score=0.964, and AUC=0.984). Fig. 6c depicts the AUC and ROC curves of the models with selected features by GWO. The METADES model achieves the highest AUC =0.984, while the MCB model achieves the lowest AUC (0.967). Fig. 7c shows the Radar Plot of the models with selected features by GWO and places the METADES Model in the outperforming category.

c) Results Analysis using the Binary Bat Algorithm (BBA) for Feature Selection: In this section, we investigate the performance of the Ensemble Dynamic Models with selected features by the BBA. Table III provides details of the results achieved by the ML models on several evaluation measures. We summarise the results as follows: Using KNOP and DESKNN with selected feature sets produced minor performance (accuracy=0.835, precision=0.897, recall=0.751, F1-score=0.818, and AUC=0.869) and (accuracy=0.836, precision=0.907, recall=0.744, F1-score=0.817, and AUC=0.871), respectively. MCB and DESP improved performance by about 0.1–0.2% above KNOP and DESKNN. KNORA-U and KNORA-E improved their performance by approximately 1.6% above KNOP and DESKNN. The highest performance was achieved with METADES (accuracy=0.915, precision=0.934, recall=0.891, F1-score=0.912, and AUC=0.959). Fig. 6b depicts the AUC and ROC curves of the models with selected features by BBA. The METADES model achieves the highest AUC of 0.959, while the DESP model achieves the lowest AUC of 0.852. Fig. 7b shows the Radar Plot of the models with selected features by BBA and places the METADES Model in the outperforming category.

d) Results Analysis using the Genetic Algorithm (GA) for Feature Selection: In this section, we investigate the performance of the Ensemble Dynamic Models with selected features by the GA. Table III provides details of the results achieved by the ML models on several evaluation measures. We summarise the results as follows: Using KNOP with selected feature sets produced minor performance (accuracy = 0.873, precision = 0.875, recall = 0.868, F1-score = 0.871, and AUC = 0.917). METADES improved their performance by approximately (1.7)% above KNOP and DESKNN, MCB and DESP improved performance by about 0.8% above METADES, and KNORA-U improved performance by about 0.22% above DESKNN, MCB, and DESP. The highest performance was achieved with KNORA-E (accuracy=0.902, precision=0.929,

recall=0.868, F1-score=0.898, and AUC=0.917). Fig. 6a depicts the AUC and ROC curves of the models with selected features by GA. The METADES model achieves the highest AUC (0.935), while the MCB model achieves the lowest AUC (0.896). Fig. 7a shows the Radar Plot of the models with selected features by GA and places the KNORA-E Model in the outperforming category.

e) Comparison Between All Models: In this paper, we investigate the performance of dynamic ensemble models utilizing all features as well as features selected using bio-inspired feature selection algorithms. As shown in Table III, a confusion matrix in Fig. 8 is used to depict and display the performance of dynamic ensemble models using GWO as a feature selection

technique and to give an overview of the model's classification errors. The METADES model with GWO as feature selection technique achieved the highest performance compared to both complete features and features selected by GA and BBA, with an accuracy of 96.5%,a precision of 97.2%, a recall of 95.7%, an F1-score of 96.4%, and an AUC of 98.4%. Conversely, the METADES model with GA-selected features demonstrated the lowest performance.

These findings emphasize the effectiveness of the approach using the METADES model and the GWO feature selection method in predicting patients at risk of developing MODS, suggesting its potential for clinical application.
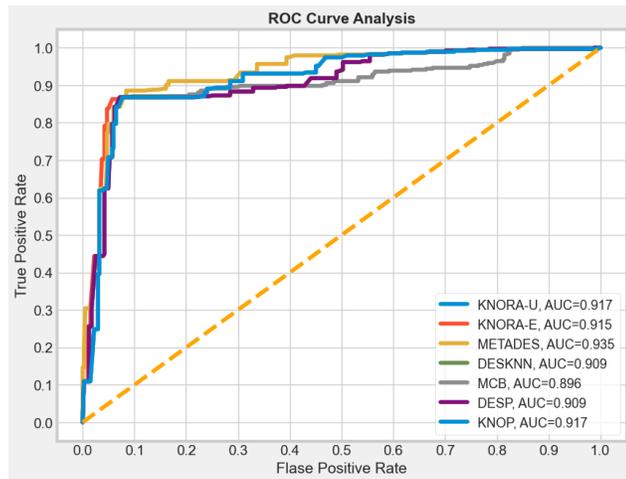
(a) ROC Curve for DES Models using GA as FS.
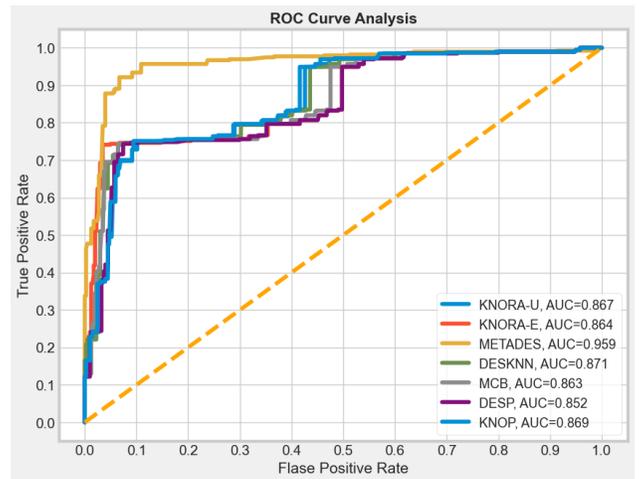
(b) ROC Curve for DES Models using BBA as FS.
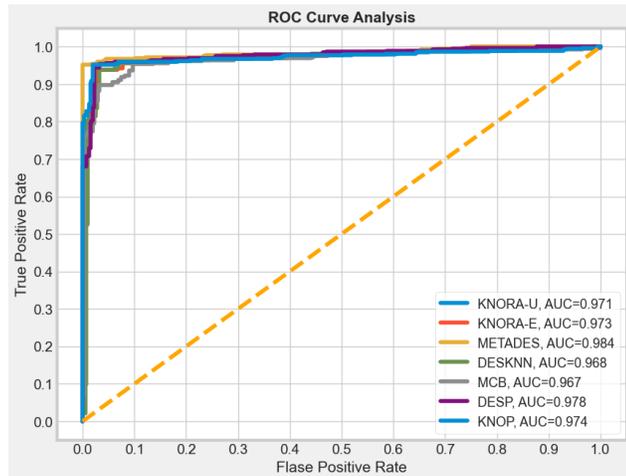
(c) ROC Curve for DES Models using GWO as FS.
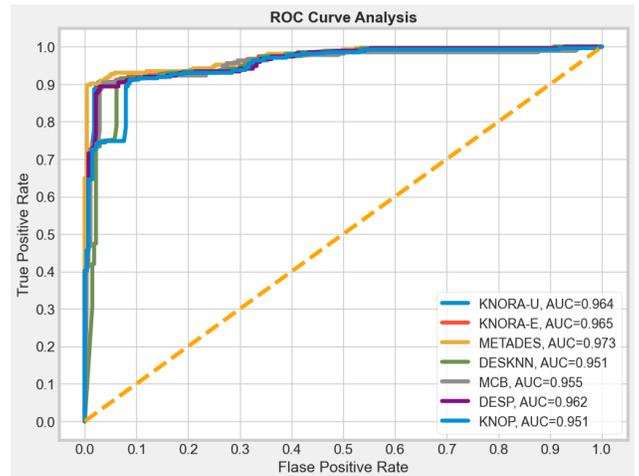
(d) ROC Curve for Dynamic Ensemble Models without Feature Selection.

Fig. 6. ROC curves of dynamic ensemble models with and without feature selection techniques.
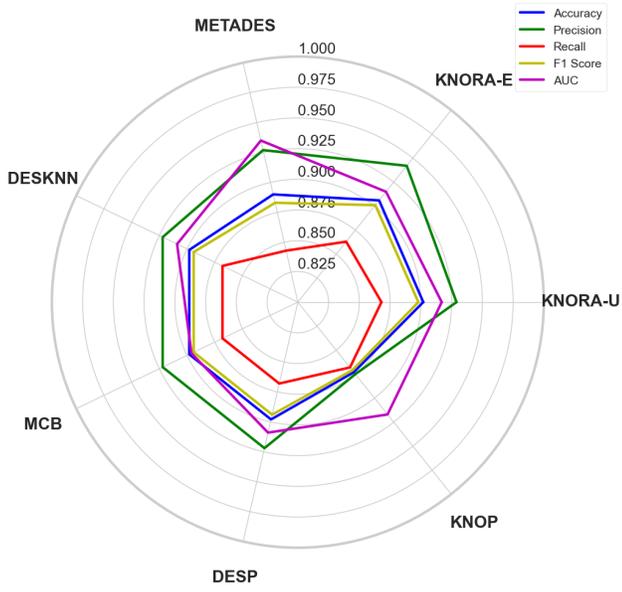
(a) Radar Plot for DES Models using GA as FS.
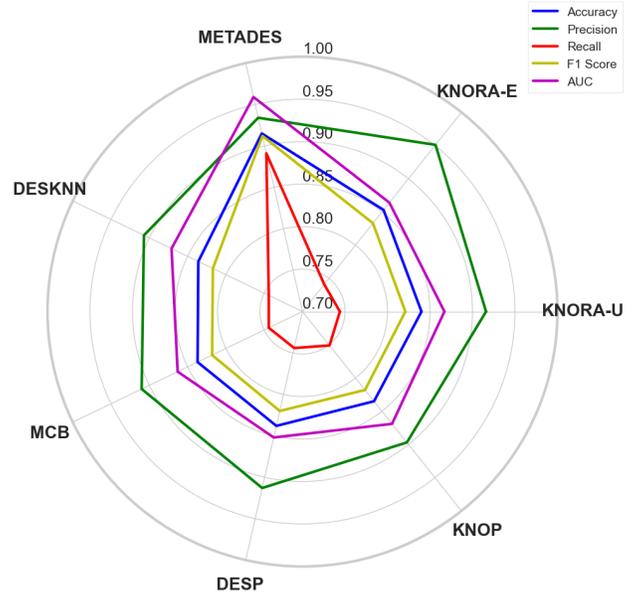
(b) Radar Plot for DES Models using BBA as FS.
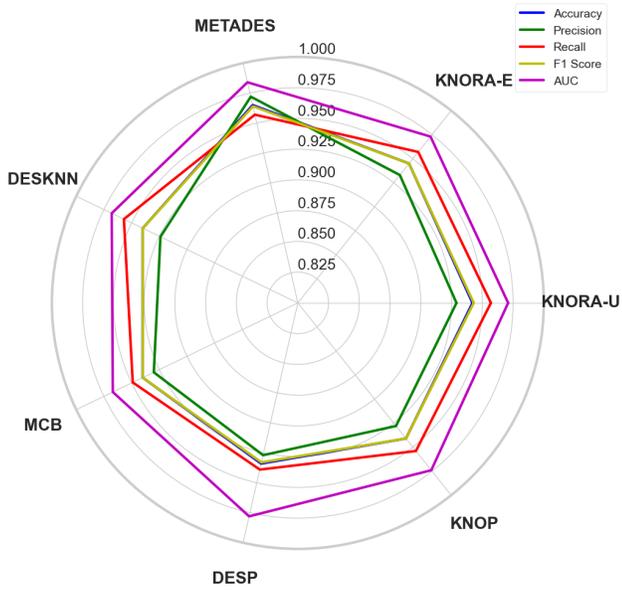
(c) Radar Plot for DES Models using GWO as FS.
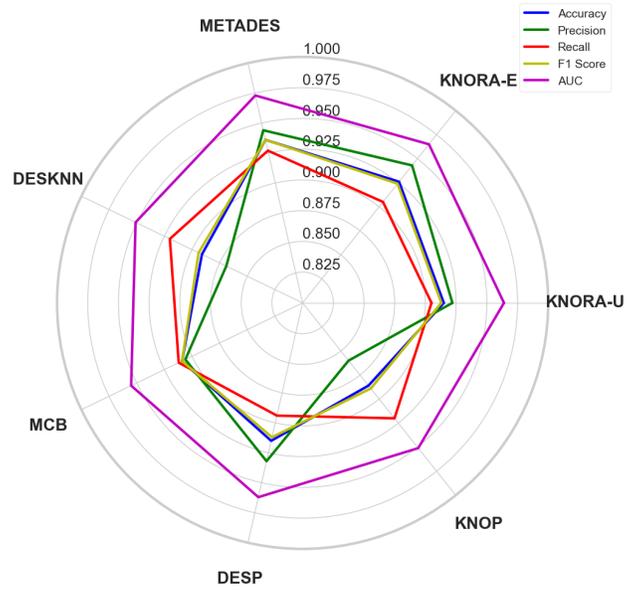
(d) Radar Plot for Dynamic Ensemble Models without Feature Selection.

Fig. 7. Radar plots of dynamic ensemble models with and without feature selection techniques.
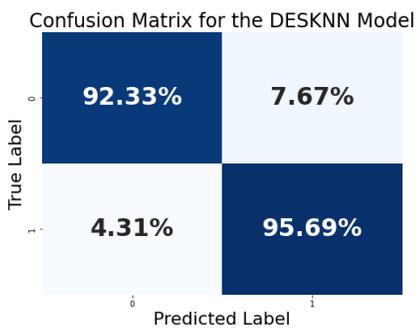
(a) Confusion Matrix for the DESKNN Model.
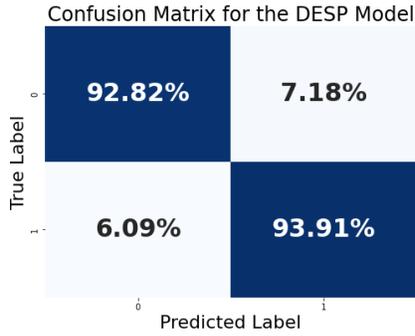
(b) Confusion Matrix for the DESP Model.
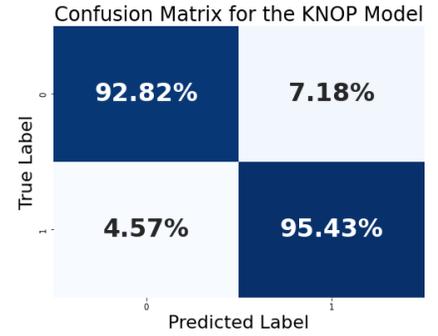
(c) Confusion Matrix for the KNOP Model.
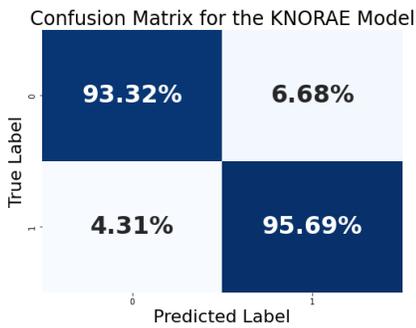
(d) Confusion Matrix for the KNORAE Model.
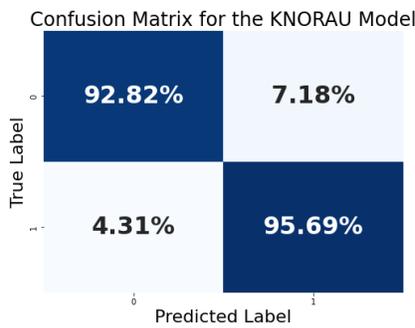
(e) Confusion Matrix for the KNORAU Model.
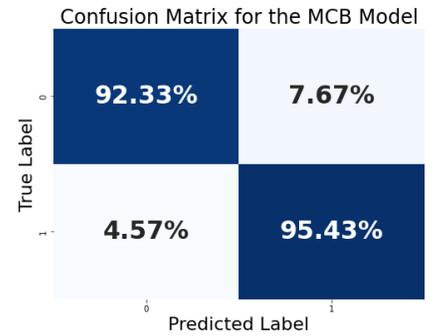
(f) Confusion Matrix for the MCB Model.
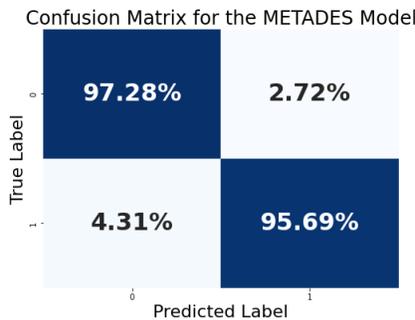
(g) Confusion Matrix for the METADES Model.

Fig. 8. Confusion matrix for dynamic ensemble models using GWO as feature selection techniques.

V. Limitations and Future Directions

Although our proposed approach is promising for the early prediction of MODS in the ICU, it has certain limitations: Firstly, the dataset used in this study includes only MIMIC III patients, specifically those admitted to the intensive care units (ICUs) of Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA. To ensure the generalizability of the model, we are planning to test it with other real-world datasets. Secondly, although dynamic ensemble models outperform deep learning models in terms of speed, the use of time-series datasets using deep learning methods such as LSTM and CNN may enhance performance. Thirdly, to be clinically accepted as a decision-support system, the approach must be interpretable. For this reason, we plan to study various methods of explainability, such as eXplainable Artificial Intelligence (XAI). Future studies will address all these limitations.

VI. Conclusion

In this work, we proposed a decision support system for the early prediction of Multi-Organ Dysfunction Syndrome (MODS) in the intensive care unit (ICU). Utilizing only non-invasive features and time-series records gathered from the initial 12 hours of admission in the ICU, the system aimed to support doctors by accelerating their decision-making process. We explored the effectiveness of dynamic ensemble selection models in predicting the risk of developing MODS within the ICU. We compared the performance of models with full features and with feature selection methods, evaluating three nature-inspired metaheuristic optimization feature selection techniques: the binary bat algorithm (BBA), grey wolf optimization (GWO), and genetic algorithm (GA) in order to select the optimal feature subset.

The proposed system was trained and evaluated on a cohort of 1,392 patients extracted from the MIMIC III dataset. The METADES model with GWO as the feature selection technique achieved the highest performance compared to models using all features or features selected by other methods. It demonstrated an accuracy of 96.5%, a precision of 97.2%, a recall of 95.7%, an F1-score of 96.4%, and an AUC of 98.4%. Conversely, the METADES model with GA-selected features exhibited the lowest performance.

These findings highlighted the effectiveness of our approach using the METADES model and the GWO feature selection method in predicting patients at risk of developing MODS, suggesting its promising potential for clinical application.

References

[1] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[2] N. M. Gourd and N. Nikitas, "Multiple organ dysfunction syndrome," *Journal of intensive care medicine*, vol. 35, no. 12, pp. 1564–1575, 2020.

[3] American Hospital Association, "Fast facts on u.s. hospitals, 2023," https://www.aha.org/statistics/fast-facts-us-hospitals, 2023, accessed on June 16, 2023.

[4] Z. Mao, C. Liu, Q. Li, Y. Cui, and F. Zhou, "Intelligent intensive care unit: Current and future trends," *Intensive Care Research*, vol. 3, no. 2, pp. 182–188, 2023.

[5] J. L. Vincent and et al., "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine." *Intensive Care Med.*, vol. 22, no. 7, pp. 707–10, 1996, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8844239.

[6] D. Johnson and I. Mayers, "Multiple organ dysfunction syndrome: a narrative review," *Canadian Journal of Anesthesia*, vol. 48, pp. 502–509, 2001.

[7] S. Y. et al., "Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome," *Crit. Care*, vol. 16, no. 6, 2012.

[8] A. Maach, J. Elalami, N. Elalami, and E. H. El Mazoudi, "An intelligent decision support ensemble voting model for coronary artery disease prediction in smart healthcare monitoring environments," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, p. 711 – 724, 2022, cited by: 0; All Open Access, Gold Open Access, Green Open Access.

[9] G. A. Alshehri and H. M. Alharbi, "Prediction of heart disease using an ensemble learning approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, p. 1089 – 1097, 2023, cited by: 0; All Open Access, Gold Open Access.

[10] S. Mamidisetti and A. M. Reddy, "A stacking-based ensemble framework for automatic depression detection using audio signals," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, p. 603 – 612, 2023, cited by: 1; All Open Access, Gold Open Access.

[11] A. Maach, J. El Alami *et al.*, "A fog-driven iot e-health framework to monitor and control asthma exacerbation," in *2019 International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2019, pp. 1–6.

[12] M. N. KP and P. Thiyagarajan, "Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: A performance analysis," *Biomedical Signal Processing and Control*, vol. 68, p. 102729, 2021.

[13] L. Malviya and S. Mal, "Cis feature selection based dynamic ensemble selection model for human stress detection from eeg signals," *Cluster Computing*, pp. 1–15, 2023.

[14] J. Wu, J. Shen, M. Xu, and M. Shao, "A novel combined dynamic ensemble selection model for imbalanced data to detect covid-19 from complete blood count," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106444, 2021.

[15] P. K. Ramtekkar, A. Pandey, and M. K. Pawar, "Accurate detection of brain tumor using optimized feature selection based on deep learning techniques," *Multimedia Tools and Applications*, pp. 1–31, 2023.

[16] A. I. Saleh and S. A. Hussien, "Disease diagnosis based on improved gray wolf optimization (igwo) and ensemble classification," *Annals of Biomedical Engineering*, vol. 51, no. 11, pp. 2579–2605, 2023.

[17] K. Chatra, V. Kuppili, D. R. Edla, and A. K. Verma, "Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function," *Medical & Biological Engineering & Computing*, vol. 57, pp. 2673–2682, 2019.

[18] P. Sharma and K. Sharma, "Fetal state health monitoring using novel enhanced binary bat algorithm," *Computers and Electrical Engineering*, vol. 101, p. 108035, 2022.

[19] F. Navazi, Y. Yuan, and N. Archer, "An examination of the hybrid meta-heuristic machine learning algorithms for early diagnosis of type ii diabetes using big data feature selection," *Healthcare Analytics*, vol. 4, p. 100227, 2023.

[20] B. Fan, J. Klatt, M. M. Moor, L. A. Daniels, L. N. Sanchez-Pinto, P. K. Agyeman, L. J. Schlapbach, and K. M. Borgwardt, "Prediction of recovery from multiple organ dysfunction syndrome in pediatric sepsis patients," *Bioinformatics*, vol. 38, no. Supplement_1, pp. i101–i108, 2022.

[21] G. Liu, J. Xu, C. Wang, M. Yu, J. Yuan, F. Tian, and G. Zhang, "A machine learning method for predicting the probability of mods using only non-invasive parameters," *Computer Methods and Programs in Biomedicine*, vol. 227, p. 107236, 2022.

[22] C. Liu, Z. Yao, P. Liu, Y. Tu, H. Chen, H. Cheng, L. Xie, and K. Xiao, "Early prediction of mods interventions in the intensive care unit using machine learning," *Journal of Big Data*, vol. 10, no. 1, pp. 1–18, 2023.

[23] T. Aşuroğlu and H. Oğul, "A deep learning approach for sepsis monitoring via severity score estimation," *Computer methods and programs in biomedicine*, vol. 198, p. 105816, 2021.

[24] A. L. Benscoter, J. A. Alten, M. R. Atreya, D. S. Cooper, J. W. Byrnes, D. P. Nelson, N. J. Ollberding, and H. R. Wong, "Biomarker-based risk model to predict persistent multiple organ dysfunctions after congenital heart surgery: a prospective observational cohort study," *Critical Care*, vol. 27, no. 1, p. 193, 2023.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[26] R. Alizadehsani and et al., "A data mining approach for diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, 2013.

[27] S. Mirjalili, S. Saremi, S. M. Mirjalili, and L. D. S. Coelho, "Multi-objective grey wolf optimizer: A novel algorithm for multi-criterion optimization," *Expert Syst. Appl.*, vol. 47, pp. 106–119, 2016.

[28] S. Mirjalili, S. M. Mirjalili, and X. S. Yang, "Binary bat algorithm," *Neural Comput. Appl.*, vol. 25, no. 3–4, pp. 663–681, 2014.

[29] J. R. Sampson, "Adaptation in natural and artificial systems (john h. holland)," *SIAM Rev.*, vol. 18, no. 3, pp. 529–530, 1976.

[30] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, and T. I. Ren, "Meta-des: A dynamic ensemble selection framework using meta-learning," *Pattern Recognit.*, vol. 48, no. 5, pp. 1925–1935, 2015.

[31] T. Woloszynski, M. Kurzynski, P. Podsiadlo, and G. W. Stachowiak, "A measure of competence based on random classification for dynamic ensemble selection," *Inf. Fusion*, vol. 13, no. 3, pp. 207–213, 2012.

[32] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Inf. Fusion*, vol. 41, pp. 195–216, 2018.

[33] A. Santana, R. G. F. Soares, A. M. P. Canuto, and M. C. P. D. Souto, "A dynamic classifier selection method to build ensembles using accuracy and diversity," in *Proc. Ninth Brazilian Symp. Neural Networks, SBRN'06*, 2006, pp. 36–41.

[34] G. Giacinto and F. Roli, "Rapid and brief communication dynamic classifier selection based on multiple classifier behaviour," vol. 34, pp. 1879–1881, 2001.

[35] P. R. Cavalin, R. Sabourin, and C. Y. Suen, "Logid: An adaptive framework combining local and global incremental learning for dynamic selection of ensembles of hmms," *Pattern Recognit.*, vol. 45, no. 9, pp. 3544–3556, 2012.