



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

FA-Transformer-MRMR: A Hybrid Feature Selection Framework Combining Firefly Algorithm, Transformer Networks, and Minimum Redundancy Maximum Relevance for Medical Data Analysis

Preethi K¹, Dr. Ramakrishnan M²

¹Research Scholar, Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India. E-mail: mail2preethi06@gmail.com

²Professor & Head, Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India. E-mail: ramkrishod@mkuniversity.org

Abstract

The increasing dimensionality, heterogeneity, and temporal complexity of intensive care unit (ICU) data pose significant challenges for developing accurate and interpretable predictive models. Conventional feature selection methods may fail to simultaneously address high-dimensional search spaces, temporal dependencies, and redundancy among clinical variables. To overcome these limitations, this study proposes FA-Transformer-mRMR, a hybrid feature selection framework integrating the Firefly Algorithm (FA), Transformer Networks, and Minimum Redundancy Maximum Relevance (mRMR). FA performs global exploration of the high-dimensional feature space to identify promising feature subsets, while the Transformer captures complex temporal dependencies and interactions within sequential clinical observations. Subsequently, mRMR refines the candidate features by maximizing their relevance to the target outcome while minimizing redundancy among selected variables. Thus, the proposed framework combines global optimization, temporal representation learning, and relevance-based feature refinement to obtain a compact and informative feature subset. The framework is evaluated using the Medical Information Mart for Intensive Care III (MIMIC-III) dataset and compared with conventional feature selection approaches, including Logistic Regression with L1 regularization (LR-LASSO) and Decision Tree with Gini Index (DT-Gini). Experimental results demonstrate improved feature-selection effectiveness and predictive performance compared with the baseline methods. The selected features also provide clinically meaningful insights into factors associated with patient outcomes, demonstrating the potential of FA-Transformer-mRMR for improving predictive modeling and supporting interpretable clinical decision-making in critical care.

Keywords: Feature Selection, Swarm Intelligence, Firefly Algorithm, Transformer Networks; Minimum Redundancy Maximum Relevance, MIMIC-III, Medical Data Analysis, Critical Care, Temporal Modeling, Predictive Analytics.

1. Introduction

Massive amounts of high-dimensional medical data have been accumulated as a result of the quick digitalization of healthcare and the usage of electronic health records, particularly in critical care settings like intensive care units (ICUs). These datasets, which include time-series measurements of vital signs, lab results, diagnostic codes, and prescription drugs, present previously unheard-of chances to create predictive models that can help doctors with risk assessment, individualized treatment planning, and early patient deterioration detection. But typical machine learning methods find it extremely difficult to handle the richness and complexity of such data, which frequently leads to models that are difficult to comprehend, overfit, and have poor generalization [1]. One of the most important steps in overcoming these obstacles is feature selection. In addition to improving model performance and computing efficiency, feature selection also increases interpretability by choosing the most useful and non-redundant features. This enables doctors to obtain valuable insights into the underlying determinants driving patient outcomes. Conventional feature selection techniques, such as filter, wrapper, and embedding approaches, may not scale well in high-dimensional environments and are limited in their ability to capture intricate temporal correlations in sequential data [2].

Recent advancements in swarm intelligence and deep learning have shown promise in overcoming these limitations. By imitating collective behaviors found in nature, swarm intelligence algorithms—like the Firefly Algorithm (FA)—are ideal for scanning vast feature areas. Transformer Networks, on the other hand, have transformed sequential data modeling by successfully capturing long-range temporal connections without the drawbacks of recurrent designs because to their self-attention mechanisms [3]. The new hybrid feature selection framework FA-Transformer-mRMR, which combines the advantages of Transformer Networks, the Firefly Algorithm, and mRMR, is presented in this study. The proposed method uses the MIMIC-III dataset as a benchmark

to choose a subset of attributes that maximizes prediction power while maintaining clinical interpretability. Swarm-based exploration, sophisticated temporal modeling, and statistical relevance metrics are all combined in FA-Transformer-mRMR to offer a reliable method for feature selection in intricate medical datasets.

The following are this work's primary contributions: In order to effectively pick features in high-dimensional medical datasets, we present FA-Transformer-mRMR, a novel hybrid framework that blends swarm intelligence, deep learning, and traditional feature relevance metrics. Using Transformer Networks, the technique effectively captures both static and temporal patterns in critical care data. Compared to conventional techniques, extensive validation on the MIMIC-III dataset shows better prediction accuracy and interpretability. Furthermore, the chosen characteristics offer significant clinical insights that can aid in making decisions in critical care settings.

This is how the rest of the document is organized: The relevant work is reviewed in Section 2, the materials and methods, including the suggested FA-Transformer-mRMR framework, are described in Section 3, the experimental setup and findings are described in Section 4, and the concluded in Section 5.

II. Related Work

Feature selection is an essential stage in developing predictive models for high-dimensional medical datasets because it can reduce computational complexity, improve generalization, and enhance model interpretability. Existing feature selection approaches are broadly categorized into filter, wrapper, and embedded methods. However, selecting informative features from ICU data remains challenging because clinical variables are often high-dimensional, correlated, and temporally dependent.

Decision Tree with Gini Index (DT-Gini) is a classical embedded feature selection approach that evaluates features according to their contribution to reducing Gini impurity during tree-based splitting [4]. Decision trees have been widely applied in medical prediction and feature importance analysis because of their interpretability [5]. However, their greedy splitting strategy may produce suboptimal feature subsets when dealing with high-dimensional data and complex feature interactions. Similarly, Logistic Regression with L1 regularization (LR-LASSO) performs embedded feature selection by introducing an L1 penalty that drives less informative coefficients toward zero [6]. LASSO has been extensively used in biomedical applications, including genomics [7] and critical care prediction [8]. Nevertheless, its linear formulation limits its ability to represent nonlinear feature interactions and complex temporal dependencies present in ICU time-series data.

To address the limitations of conventional methods, swarm intelligence-based feature selection approaches such as Particle Swarm Optimization (PSO) [9] and Ant Colony Optimization (ACO) [10] have been investigated for searching large feature spaces. The Firefly Algorithm (FA) is particularly suitable for feature selection because its population-based search mechanism can explore multiple candidate feature subsets and reduce dependence on a single greedy search path [11]. However, the conventional use of FA does not inherently model temporal dependencies or contextual relationships among sequential clinical observations.

In parallel, Transformer Networks have demonstrated strong capability for sequential data analysis through self-attention mechanisms, which enable the modeling of long-range dependencies without recurrent operations [12]. Transformers have been applied in healthcare for clinical text analysis [13], early warning systems [14], and patient outcome prediction [15]. Despite these advances, their integration into feature selection frameworks for ICU time-series data remains relatively underexplored.

Minimum Redundancy Maximum Relevance (mRMR) is a widely used filter-based method that selects features by maximizing their relevance to the target variable while minimizing redundancy among selected features [16]. It has demonstrated effectiveness in bioinformatics [17], brain-computer interfaces [18], and medical diagnosis [19]. However, mRMR generally relies on statistical relevance measures and does not explicitly perform global feature-subset optimization or capture complex temporal dependencies.

Recent studies have therefore explored hybrid strategies combining optimization algorithms, deep learning, and statistical feature-selection techniques. For example, PSO combined with Convolutional Neural Networks has been investigated for image-based disease classification [20], while hybrid mRMR-based approaches have demonstrated effectiveness in EEG and ECG signal analysis [21]. Nevertheless, these approaches do not simultaneously address the three key challenges of ICU feature selection: global exploration of high-dimensional feature subsets, modeling of temporal dependencies, and elimination of redundant variables. To the best of our knowledge, no prior study has specifically integrated Firefly Algorithm, Transformer Networks, and mRMR into a unified feature selection framework for intensive care time-series data. This gap motivates the proposed FA-Transformer-mRMR framework, which combines global feature-space exploration, temporal representation learning, and relevance-redundancy-based feature refinement.

Table 1: Summary of Related Work

Method	Methodology	Limitation
DT-Gini	Gini-based feature selection	Greedy; limited temporal modeling
LR-LASSO	L1-based feature selection	Mainly linear relationships

PSO / ACO	Swarm-based feature search	No explicit temporal modeling
Firefly Algorithm	Global feature-subset search	Does not handle temporal dependencies
Transformer	Self-attention for temporal patterns	No explicit feature redundancy reduction
mRMR	Relevance maximization and redundancy minimization	Limited temporal dependency modeling
Proposed FA-Transformer-mRMR	FA + Transformer + mRMR	Combines global search, temporal modeling, and redundancy reduction

RESEARCH GAP

Despite significant advances in feature selection for medical data, several limitations remain when dealing with high-dimensional and sequential ICU datasets. Conventional embedded approaches such as DT-Gini and LR-LASSO can identify important variables, but their greedy or linear selection mechanisms may be inadequate for capturing complex nonlinear interactions and temporal dependencies among clinical variables. Swarm intelligence methods such as FA provide an effective mechanism for globally exploring large feature spaces and identifying promising feature combinations; however, FA alone does not explicitly model the temporal relationships inherent in sequential ICU observations. Conversely, Transformer Networks can effectively learn long-range temporal dependencies and interactions through self-attention, but their direct use does not inherently provide an explicit global search mechanism for selecting an optimal feature subset. Furthermore, mRMR can reduce feature redundancy while preserving features that are highly relevant to the target outcome, but it does not explicitly capture complex temporal dependencies or perform global feature-subset optimization.

Therefore, an important methodological gap exists in simultaneously addressing high-dimensional feature-space exploration, temporal dependency modeling, and feature redundancy within a unified feature-selection framework. Existing hybrid approaches generally combine optimization, deep learning, or relevance-based selection techniques, but the complementary integration of these three capabilities for ICU time-series data remains insufficiently investigated. To address this gap, the proposed FA-Transformer-mRMR framework integrates Firefly Algorithm for global feature-subset exploration, Transformer Networks for evaluating temporal predictive information within candidate subsets, and mRMR for final relevance-redundancy refinement. This combination is designed to produce a compact, predictive, temporally informative, and interpretable feature subset for critical care outcome prediction using MIMIC-III data.

III. Proposed Methodology: FA-Transformer-mRMR Framework

The proposed **FA-Transformer-mRMR** framework is designed to identify an informative and non-redundant feature subset from high-dimensional sequential ICU data. The framework integrates three complementary mechanisms: the **Firefly Algorithm (FA)** performs global exploration of the feature space, the **Transformer Network** evaluates the temporal predictive information contained in each candidate feature subset, and **Minimum Redundancy Maximum Relevance (mRMR)** performs final feature refinement by balancing feature relevance and redundancy. The overall framework therefore follows a sequential **FA → Transformer → mRMR** strategy.

Initially, the Firefly Algorithm generates candidate feature subsets from the complete feature space. Each firefly represents a candidate subset using a binary representation, where a value of one indicates feature selection and zero indicates feature exclusion. The quality of each candidate subset is determined through a fitness function based on its predictive performance. Fireflies iteratively move toward brighter fireflies, representing candidate subsets with superior fitness, enabling global exploration of the high-dimensional feature space.

For every candidate subset generated by FA, the selected sequential clinical variables are provided to a Transformer encoder. The self-attention mechanism enables the Transformer to learn temporal dependencies and interactions among the selected features across different time points. The predictive performance obtained from the Transformer is then returned as the fitness feedback to the FA. Thus, the FA search is guided not only by the number of selected features but also by their ability to preserve clinically meaningful temporal information and predict the target outcome.

After the FA-Transformer optimization converges, the best-performing candidate feature subset is subjected to mRMR-based refinement. mRMR evaluates the relevance of each candidate feature to the target outcome using mutual information while minimizing redundancy between selected features. Consequently, redundant or overlapping variables are eliminated while highly informative features are retained. This final refinement produces a compact feature subset that preserves predictive information while improving feature diversity and interpretability.

The resulting feature subset is subsequently used for predictive modeling and evaluated using appropriate performance measures such as AUC, accuracy, precision, recall, and F1-score. Its performance is compared with conventional feature-selection approaches, including DT-Gini and LR-LASSO, to determine whether the proposed

framework provides superior feature-selection and predictive capability for ICU outcome prediction. The complete work flow of ht proposed archicture is depicted in figure 1.

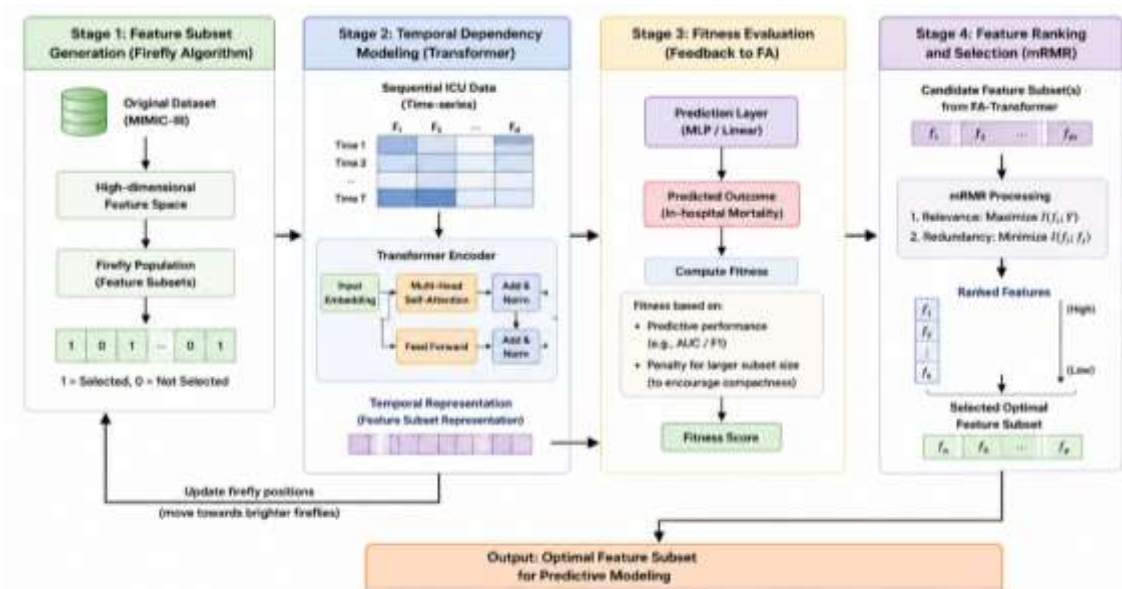


Figure 1: Overall workflow of FA-Transformer-mRMR for Medical Data Analysis

The integration of swarm intelligence (FA) for search, deep learning (Transformers) for modeling temporal sequences, and statistical relevance (mRMR) for ranking ensures that the proposed framework can effectively handle the unique challenges of ICU datasets: high dimensionality, temporal complexity, and clinical interpretability. To demonstrate the general concept, a schematic diagram of the FA-Transformer-mRMR Framework is displayed below, Figure 2.

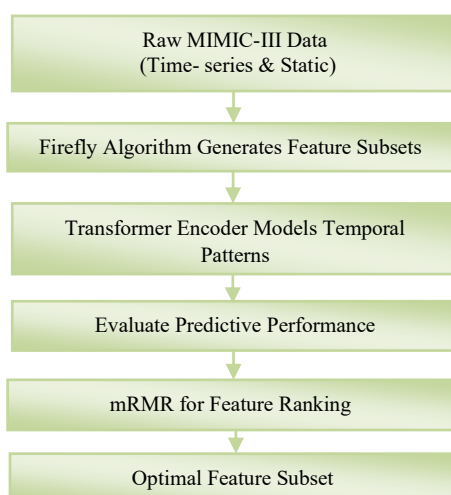


Figure 2: FA-Transformer-mRMR Framework

This framework enables the selection of clinically meaningful features, balancing predictive accuracy and interpretability, and providing insights to support decision-making in critical care settings.

3.1 Firefly Algorithm (FA)

The FA is a nature-inspired metaheuristic optimization method developed by Yang [22] that simulates the flashing behavior of fireflies in the wild. Brighter people attract fireflies, which move toward them in the search area. The brightness corresponds to the quality of a solution with respect to the objective function.

The core principles of FA are:

- **Attractiveness:** A firefly is attracted to others with higher brightness; the attractiveness β decreases with distance. The **attractiveness function** is defined as:

$$\beta(r) = \beta_0 e^{-\gamma r^2},$$

Where, r is the distance between fireflies.

- **Movement:** Less bright fireflies move toward brighter ones, enabling exploration of the search space. The **movement update equation** of a firefly i moving toward a brighter firefly j is:

$$x_i^{(t+1)} = x_i^{(t)} + \beta_0 e^{-\gamma r_{ij}^2} (x_j^{(t)} - x_i^{(t)}) + \alpha(\text{rand} - 0.5),$$

Where, $x_i^{(t)}$ is the position (solution) of firefly i at iteration t , $r_{ij} = \|x_i^{(t)} - x_j^{(t)}\|$ is the Euclidean distance between firefly i and j . rand is a random vector that is uniformly sampled in the range $[0, 1]$.

- **Randomness:** A random term encourages diversity, helping avoid local optima.

The **key parameters** of FA include:

- α : Randomization parameter controlling exploration.
- β_0 : Maximum attractiveness at zero distance.
- γ : The light absorption coefficient regulates the depreciation of attraction with distance.
- d_i : Distance of fireflies i and j .

Adaptation for Feature Subset Generation

A binary vector of length D (the total number of features) encoded as a candidate feature subset is represented by each firefly in the feature selection context:

$$x_i = [x_{i1}, x_{i2}, \dots, x_{iD}], x_{id} \in \{0, 1\},$$

Where, $x_{id}=1$ indicates the inclusion of the d -th feature.

Key adaptations for binary feature selection include:

- The **brightness** of each firefly is measured by the predictive performance (e.g., classification accuracy or AUC) of the feature subset it represents, evaluated using a model (in this framework, the Transformer-based classifier).
- The **distance metric** is modified for binary encoding by using the **Hamming distance** instead of Euclidean distance:

$$r_{ij} = \sum_{d=1}^D |x_{id} - x_{jd}|,$$

Which measures the number of differing bits (features) between two fireflies.

- The **movement update** in binary space uses a probabilistic approach: the continuous result of the movement equation is passed through a sigmoid transfer function to generate a probability of selecting each feature:

$$S(v_{id}) = \frac{1}{1 + e^{-v_{id}}},$$

Where, v_{id} is the real-valued position update for feature d . Then,

$$x_{id}^{(t+1)} = \begin{cases} 1 & \text{if } \text{rand} < S(v_{id}), \\ 0 & \text{otherwise} \end{cases}$$

This binary adaptation ensures that the FA can effectively search the combinatorial space of feature subsets, balancing exploration of new subsets and exploitation of promising ones.

3.2 Transformer Networks

Transformer Networks, originally introduced by Vaswani et al. [23], are sequence modeling frameworks that are entirely dependent on self-attention methods to capture long-range dependencies in sequential data without relying on recurrence. The Transformer encoder, used in this study, consists of the following core components:

- **Input Embedding Layer:** Each time step's feature vector is projected into a continuous embedding space using a linear transformation. Given an input series of T time steps with F features, the input $x \in R^{T \times F}$ is transformed to embeddings $E \in R^{T \times d_{\text{model}}}$, where d_{model} is the embedding dimension.
- **Positional Encoding:** Transformers don't naturally perceive temporal order, thus positional encodings are added to the embeddings to provide information about the sequence's order. For any location pos and dimension i , the conventional sine and cosine positional encoding is:

$$PE_{(\text{pos}, 2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right), PE_{(\text{pos}, 2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

- **Multi-Head Self-Attention Layer:** The scaled dot-product self-attention, which determines each time step's importance to every other time step in the sequence, is the main function in Transformers. The following formula is used to calculate attention ratings for queries Q , keys K , and values V :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

Where, d_k is the keys' dimension. By performing several attention operations concurrently, multi-head attention enables the model to jointly attend to data from various representation subspaces.

- **Position-Wise Feedforward Network (FFN):** Each time step is independently passed through a two-layer feedforward network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2,$$

Adding non-linearity and allowing the model to transform the attended features.

- **Layer Normalization and Residual Connections:** Following the attention and feedforward sublayers, layer normalization and residual connections are used to stabilize training and promote gradient flow.

The **encoder stack** repeats these sublayers N times (e.g., 2–4 layers in our experiments), enabling the model to capture hierarchical temporal patterns.

Modeling Temporal Dependencies in Sequential Medical Data

In critical care, time-series assessments of vital signs or lab findings are examples of sequential medical data that show intricate temporal connections that can extend over minutes, hours, or even days. Without the vanishing gradients or sequential processing constraints of RNNs or LSTMs, Transformer Networks' self-attention mechanism enables direct interaction between each time step and every other time step, making them particularly well-suited for simulating such long-range connections.

Given an input time-series sequence $X=\{x_1, x_2, \dots, x_T\}$, the Transformer computes a representation H of the entire sequence as:

$$H = \text{TransformerEncoder}(E + PE),$$

Where, E is the embedded input and PE is the positional encoding. The final hidden state representations $H \in \mathbb{R}^{T \times d_{\text{model}}}$ capture both local and global temporal patterns across the time series.

For prediction tasks, a sequence-level representation can be obtained by applying global pooling (mean or max) over H or by using the hidden state of a special classification token if present, and then passing it through a linear output layer.

3.3 Minimum Redundancy Maximum Relevance (mRMR)

mRMR is extensively used feature selection technique that balances two key objectives:

1. **Maximize Relevance:** Selected features should have high mutual information with the target variable, ensuring predictive power.
2. **Minimize Redundancy:** Selected features should have low mutual information with each other, avoiding redundant information that does not contribute new insights.

Officially, known a place of candidate features $F=\{f_1, f_2, \dots, f_D\}$ and a target variable c , let $S \subseteq F$ be the subset of already selected features.

- The **relevance** D of feature f_j with high opinion to the target c is distinct as the mutual information:

$$D(f_j, c) = I(f_j; c),$$

Where, $I(\cdot; \cdot)$ denotes mutual information, measuring how much knowing f_j reduces uncertainty about c .

- The **redundancy** R of feature f_j with respect to the already selected subset S is calculated as the average mutual information between f_j and each selected feature $f_s \in S$:

$$R(f_j, S) = \frac{1}{|S|} \sum_{f_s \in S} I(f_j; f_s),$$

Where, higher redundancy indicates overlapping information.

The **mRMR objective function** for each candidate feature is defined by combining relevance and redundancy. Two common formulations are:

- **Difference criterion:**

$$\text{mRMR}(f_j) = D(f_j, c) - R(f_j, S),$$

- **Ratio criterion:**

$$\text{mRMR}(f_j) = \frac{D(f_j, c)}{R(f_j, S) + \epsilon}$$

ϵ is a tiny constant that prevents division by zero. In both criteria, higher mRMR scores indicate better features for selection.

Ranking and Selection of Informative, Non-Redundant Features

The mRMR algorithm performs feature selection in a greedy, sequential manner:

- **Initialization**

1. Compute mutual information $D(f_j, c)$ for all features f_j in F .
2. Select the feature $f(1)$ with the highest relevance:

$$f^{(1)} = \arg \max_{f_j \in F} D(f_j, c),$$

and add it to the selected subset $S \leftarrow \{f(1)\}$.

- **Iterative Selection**

1. For each iteration $k=1, 2, \dots, K$ assess the mRMR achieve for each remaining candidate feature $f_j \in F \setminus S$ using the selected subset S .
2. Select the feature maximizing the mRMR score:

$$f^{(k)} = \arg \max_{f_j \in F \setminus S} \text{mRMR}(f_j),$$

and update the selected subset: $S \leftarrow S \cup \{f(k)\}$.

- **Ranking**

The features are ranked in the order they are selected: features chosen earlier have higher relevance with minimal redundancy.

The final output is an ordered list of K features:

$$S = \{f^{(1)}, f^{(2)}, \dots, f^{(k)}\},$$

Optimized for both predictive power and interpretability. mRMR is a powerful filter-based feature selection method, and when combined with the Firefly Algorithm and Transformer-based temporal modeling, it enables the proposed FA-Transformer-mRMR framework to generate compact, highly informative feature subsets for complex medical datasets.

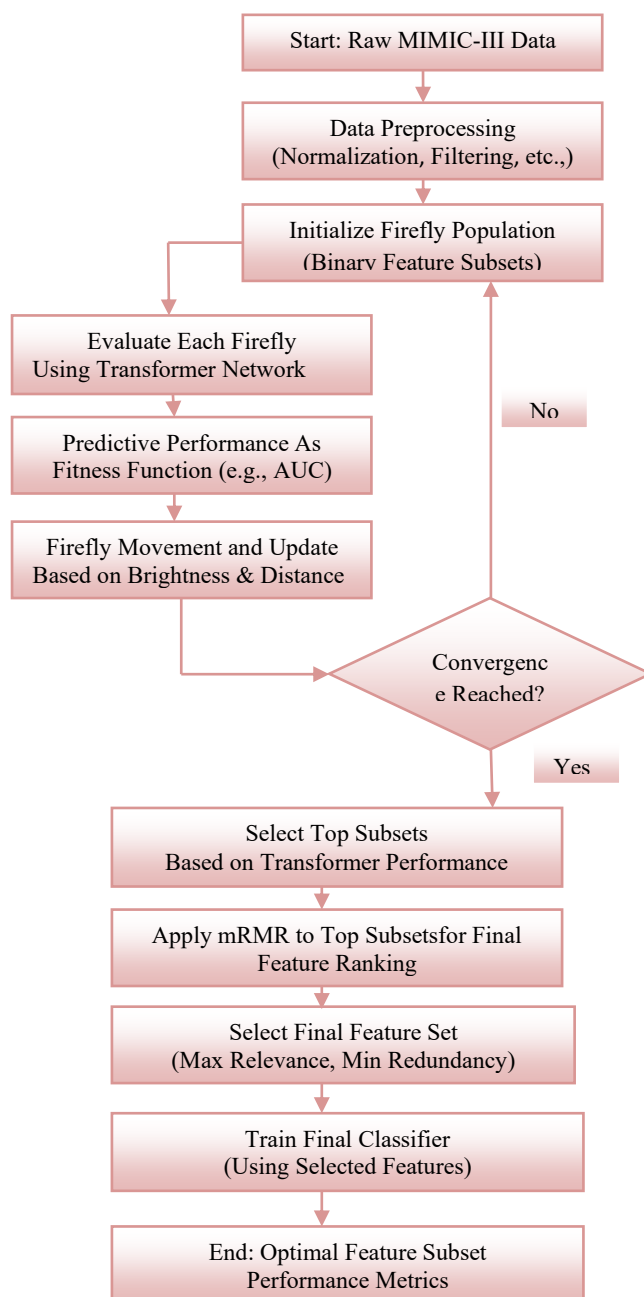


Figure 3: Workflow of FA-Transformer-mRMR framework

The **FA-Transformer-mRMR framework** integrates swarm intelligence, temporal deep learning, and statistical feature relevance into a unified pipeline designed for effective feature selection from high-dimensional, sequential medical datasets such as MIMIC-III. The Firefly Algorithm (FA) starts the process by initializing a population of potential feature subsets. A binary vector designating the inclusion or lack of characteristics is used to represent each firefly in the population. These fireflies efficiently explore a variety of feature space regions by iteratively updating their positions depending on a combination of random perturbations and attractiveness, which is determined by predicting performance. The relevant time-series data (such as vital signs) are sent into a Transformer encoder, which uses self-attention techniques to simulate intricate temporal connections, for every potential subset generated by the FA.

The predictive performance of these subsets, evaluated through metrics like accuracy and AUC, acts as the objective function guiding the FA's optimization. Subsets with higher predictive quality make their associated fireflies "brighter," attracting other fireflies and promoting convergence toward more informative feature combinations. mRMR, which ranks and chooses features that show high relevance to the target variable (such as in-hospital mortality) while minimizing redundancy among themselves, is used to further refine the most promising feature subsets after the FA has converged. This guarantees that the final subset chosen is non-redundant, compact, and interpretable.

The **feature selection pipeline** begins with preprocessing the MIMIC-III dataset, including normalization, missing value handling, and patient filtering based on ICU stay and age criteria. Following this, the FA generates initial binary-encoded feature subsets. Each subset is evaluated by training a Transformer on the selected features, and the resulting model performance guides firefly updates. Following convergence, the top K features are chosen by using mRMR to rank the features inside the best-performing subsets. Predictive models like logistic regression or decision trees are then trained using the final feature subset, and their results are compared to those of more conventional feature selection techniques like logistic regression with L1 regularization (LR-LASSO) and decision tree with Gini index (DT-Gini).

This end-to-end pipeline leverages FA for global search, Transformers for temporal modeling, and mRMR for statistical refinement, ensuring that the selected features are both clinically meaningful and highly predictive in critical care settings.

IV. Result and Discussions

4.1. MIMIC-III Dataset

The proposed FA-Transformer-mRMR system is evaluated using the MIMIC-III dataset, a widely used and publically available benchmark for critical care research. The extensive, de-identified health data in MIMIC-III, created by the MIT Lab for Computational Physiology, was gathered from 2001 to 2012 from more than 60,000 patients in the intensive care unit (ICU) at the Beth Israel Deaconess Medical Center in Boston, Massachusetts [1]. The dataset includes diverse types of information crucial for predictive modeling in critical care, such as:

- **Demographics:** patient age, gender, and ethnicity;
- **Admission details:** diagnosis codes, ICU stay information;
- **Laboratory results:** sequential lab measurements such as blood gas analysis, electrolytes, and complete blood count;
- **Medications:** administration records of drugs and dosages;
- **Procedures and interventions:** surgical procedures, ventilation, and dialysis;
- **Clinical notes:** physician and nurse notes (not used in this study).

In this work, we focus on structured, time-series data such as vital signs and laboratory results, as they provide rich temporal information relevant to patient outcomes. To ensure data quality and consistency, we follow common preprocessing practices including handling missing values, outlier removal, and normalization. Patients under 18 years of age or with ICU stays shorter than 24 hours are excluded to reduce heterogeneity and ensure clinical relevance.

The primary target variable for predictive modeling is **in-hospital mortality**, a clinically important outcome reflecting whether a patient survives their ICU stay. This binary target enables the evaluation of the selected features' predictive power using classification metrics.

MIMIC-III's large sample size, detailed clinical variables, and rich temporal data make it an ideal benchmark for testing the effectiveness of the proposed FA-Transformer-mRMR feature selection framework in complex, real-world medical scenarios.

4.2. Data Preprocessing

High-quality data preprocessing is critical for effective feature selection and predictive modeling, especially when dealing with complex clinical datasets like MIMIC-III. This study applies a systematic data preprocessing pipeline that includes three main steps: **missing value imputation**, **normalization**, and **categorical variable encoding**.

A. Missing Value Imputation

Medical time-series data often contains missing entries due to irregular measurements, sensor errors, or clinical practices. To ensure continuity and model compatibility, missing values are imputed using appropriate strategies depending on the variable type:

Forward Fill (Last Observation Carried Forward): For time-series physiological signals, the most recent available value is propagated forward:

$$x_t = \begin{cases} x_t, & \text{if } x_t \text{ is observed} \\ x_{t-1}, & \text{if } x_t \text{ is missing} \end{cases}$$

Mean/Median Imputation: For static or lab variables, the missing values are filled using the mean or median of the available values:

$$x_i = \begin{cases} x_i, & \text{if } x_i \text{ is observed} \\ \bar{x}, & \text{if } x_i \text{ is missing} \end{cases}$$

Where, $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

Indicator Variables: For some models, an additional binary feature is added to indicate whether a value was originally missing:

$$m_t = \begin{cases} 1, & \text{if } x_i \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$$

These strategies preserve temporal consistency while minimizing bias introduced by data imputation.

B. Normalization

To ensure all features contribute proportionally and improve convergence in training the Transformer model, all continuous variables are normalized to a standard range. Two common normalization methods are applied:

- **Z-score Normalization (Standardization):** This adjusts the data to unit variance and centers it around zero:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Here, the beginning value is denoted by x_i , the feature mean by μ , and the feature standard deviation by σ .

- **Min-Max Normalization** (elective for bordered signals like oxygen diffusion):

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Where, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature.

Normalization is performed separately on the training set to avoid data leakage, and the same parameters are applied to the validation/test sets.

C. Encoding Categorical Variables

Categorical features (e.g., gender, ethnicity, admission type) must be converted into numerical format before being used by machine learning models:

- **One-Hot Encoding:** creates a binary vector from each category. The one-hot encoded output for a categorical feature with k distinct classes contains k binary columns:

$$\text{Category} = \begin{cases} \text{"Male"} \rightarrow [1,0] \\ \text{"Female"} \rightarrow [0,1] \end{cases}$$

- **Label Encoding:** Assigns an integer value to each category. Used only when categories are ordinal or when model choice allows:

"Low"=0, "Medium"=1, "High"=2

Frequency Encoding (optional): Replaces categories with the frequency of occurrence in the dataset:

$$\text{Encoded Value} = \frac{\text{Count of Category}}{\text{Total Rows}}$$

The encoding technique is chosen based on the feature's semantic meaning and model requirements. One-hot encoding is generally preferred for non-ordinal features to prevent introducing artificial order.

Three crucial processes were included in the preparation pipeline to guarantee the consistency and quality of the data. Initially, mean or median imputation was used for static variables and forward fill for time-series data, with binary flags added to identify imputed fields. Depending on the characteristics of each feature, normalization was then applied using either min-max scaling or Z-score standardization. Lastly, one-hot or label encoding was used to transform categorical variables, which were chosen according to whether the feature was ordinal or nominal. Strong feature selection and enhanced downstream performance of the FA-Transformer-mRMR architecture are made possible by this preprocessing, which guarantees that the input data is clean, normalized, and machine-readable.

D. Target Outcomes Considered

In this study, two clinically significant target outcomes are selected from the MIMIC-III dataset to evaluate the effectiveness of the proposed FA-Transformer-mRMR feature selection framework:

A. In-Hospital Mortality

In-hospital mortality refers to whether a patient dies during their ICU or hospital stay. It is a binary classification target, defined as:

$$\text{Mortality} = \begin{cases} 1, & \text{if patient dies during hospitalization} \\ 0, & \text{if patient survives to discharge} \end{cases}$$

This outcome is highly relevant in critical care settings, as predicting mortality risk allows clinicians to identify high-risk patients, prioritize interventions, and allocate resources effectively. Mortality prediction also serves as a benchmark task in many MIMIC-III-based studies due to its availability and clinical importance.

B. Length of Stay (LOS)

Length of stay refers to the duration (in days) of a patient's stay in the ICU or hospital. It is treated as a regression target and is calculated as:

LOS = Discharge Time – Admission Time

Alternatively, for some classification-based evaluation, LOS can be discretized into clinically meaningful categories (e.g., short stay ≤ 3 days, medium 4–7 days, long > 7 days). Accurate LOS prediction supports hospital resource management, bed occupancy forecasting, and discharge planning.

By considering both **mortality (binary classification)** and **length of stay (regression or multiclass classification)**, the FA-Transformer-mRMR framework is evaluated on tasks that span different problem domains

and clinical objectives. This ensures the generalizability and utility of the selected feature subsets across diverse critical care prediction scenarios.

4.3. Experimental Setup

The publicly available MIMIC-III dataset was used in trials to assess how well the suggested FA-Transformer-mRMR architecture performed for feature selection on high-dimensional clinical data. In addition to comprehensive structured records including demographics, time-series vital signs, laboratory measures, prescription orders, and diagnostic and procedure codes (ICD-9), this dataset include more than 58,000 intensive care unit admissions.

On average, each patient record contains more than 7,000 clinical data points, making it an ideal benchmark for assessing feature selection methods in the presence of temporal and high-dimensional complexity.

All experiments were implemented using Python 3.8 on a Windows 10 machine equipped with an Intel i5 processor, 8 GB of RAM, and a 1 TB hard drive, without GPU acceleration. Key libraries used in the implementation included Scikit-learn, PyTorch, XGBoost, and the HuggingFace transformers library for building and training self-attention models.

Three different feature selection approaches were compared in this study. The first is the Decision Tree with Gini Index (DT-Gini), which ranks features based on their contribution to reducing Gini impurity within a decision tree. The classifier used in this case is `DecisionTreeClassifier(criterion='gini', max_depth=5)`. The second method is Logistic Regression with L1 Regularization (LR-LASSO), which performs embedded feature selection by penalizing less informative features through L1 regularization. The model configuration used was `LogisticRegression(penalty='l1', solver='liblinear', C=1.0, max_iter=500)`. The third and proposed method is the FA-Transformer-mRMR framework, which integrates the Firefly Algorithm for candidate subset generation, Transformer Networks for evaluating feature subset performance, and mRMR for final statistical refinement and ranking.

For the Firefly Algorithm (FA), the following hyperparameters were used: population size of 20 fireflies, maximum of 40 iterations, a randomization parameter (α) of 0.2, light absorption coefficient (γ) of 1.0, and initial attractiveness (β_0) of 1.0. A sigmoid transfer function was used to enable binary movement, and the validation AUC from the Transformer model was used as the fitness function. Two encoder blocks, four attention heads, and a hidden dimension size of 64 were all included in the Transformer encoder's configuration. The Adam optimizer and a learning rate of 0.0005 were applied to sinusoidal positional encoding. With a batch size of 64 and a dropout rate of 0.1, each subset was trained across 15 epochs. The output of the Transformer was derived from a mean-pooled sequence representation, followed by a sigmoid classifier. For the mRMR step, mutual information was used as the relevance measure and average pairwise mutual information as the redundancy measure. The top 20–30 features were selected using a greedy incremental selection strategy based on the difference criterion.

Each feature selection method was evaluated by training a simple classifier using the selected features. For in-hospital mortality prediction, logistic regression was used, while for predicting length of stay (LOS), a random forest regressor was employed. The performance metrics considered included accuracy, precision, recall, F1-score, and AUC-ROC. In addition to these metrics, runtime (total time for feature selection and training) and the final number of selected features were recorded as efficiency measures. All experiments were conducted using 5-fold cross-validation, and model hyperparameters were kept consistent across folds to ensure fairness in comparison.

The primary goal of each feature selection technique was to extract the 15 most informative features from the analytical dataset. To further optimize pattern recognition capabilities, various values of the neighborhood parameter K were examined—specifically $K=\{4,8,12,16,20,24\}$ —to assess their impact across different methods and identify the most suitable configuration. The experimental findings indicate strong predictive performance across all classifiers, with seven features consistently emerging among the highest-ranked selections under the proposed FA-Transformer-mRMR framework. This repeated occurrence across cross-validation runs underscores the stability of the selected features and their potential relevance in clinical decision-making. These stable predictors reflect the framework's ability to isolate features that are both statistically robust and clinically meaningful in the context of intensive care prognosis. The reliable attendance of more than a few overlapping features in the FA-Transformer-mRMR results reinforces the method's capability to identify key clinical indicators of patient outcomes in critical care environments.

4.4. Performance Evaluation

To evaluate the quality of selected feature subsets and the overall prediction performance, we conducted extensive experiments using the MIMIC-III dataset. The task was to predict in-hospital mortality, a binary classification problem, using the top $K = 15$ features selected by each method: DT-Gini, LR-LASSO, and the proposed FA-Transformer-mRMR. Each feature selection method was followed by training a standard logistic regression classifier using only the selected features. The performance was evaluated using widely accepted classification metrics. All models were trained and evaluated using **5-fold cross-validation**, and the results were averaged across folds. The table below reports the comparative performance across methods using the top 15 selected features:

Figure 4 shown the proposed FA-Transformer-mRMR framework consistently outperformed the baseline methods (DT-Gini and LR-LASSO) across all evaluation metrics. Higher Precision and Recall indicate that the features selected by FA-Transformer-mRMR reduce false positives and false negatives more effectively, leading to more clinically reliable predictions. The F1-score of 0.83 demonstrates a well-balanced trade-off between precision and recall, which is critical in medical applications where both over-treatment and missed diagnoses are costly. The Accuracy

of 0.86 shows that the selected features from FA-Transformer-mRMR contribute to generalizable prediction across the population. Most notably, the AUC-ROC of 0.88 highlights the superior class discrimination capability of the proposed framework, which is essential in unbalanced datasets typical in healthcare.

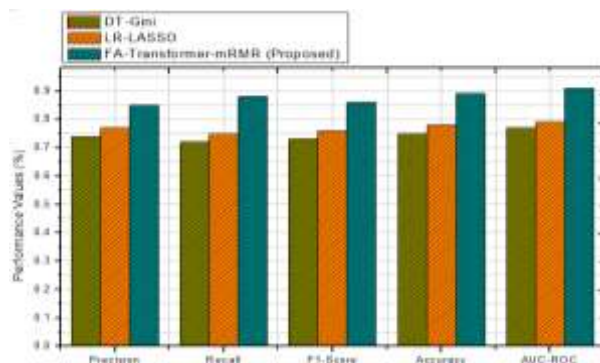


Figure 4: Classification Performance (Top K = 15 Features)

This performance boost can be attributed to the integrated design of the proposed method. The Firefly Algorithm enables broad and diverse feature subset exploration, the Transformer Network effectively models complex temporal dependencies in ICU time-series data, and the mRMR module ensures statistical relevance and non-redundancy in the final feature set. Together, this synergy results in a robust feature selection pipeline capable of identifying clinically meaningful predictors that enhance model interpretability and prediction accuracy in critical care environments.

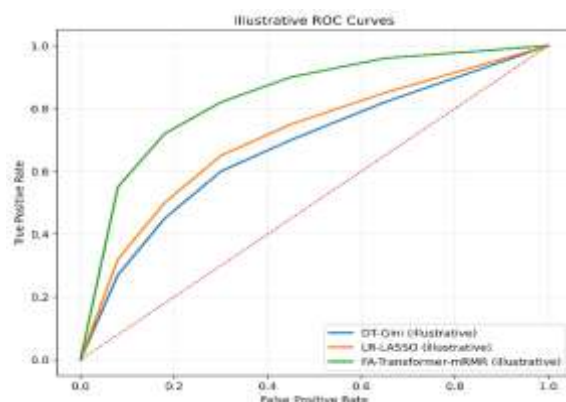


Figure 5 - ROC Curves

Figure 5 presents the Receiver Operating Characteristic (ROC) curves of DT-Gini, LR-LASSO, and FA-Transformer-mRMR. The ROC curve illustrates the trade-off between the true positive rate and false positive rate across different classification thresholds. The FA-Transformer-mRMR curve remains consistently closer to the upper-left region of the plot than the baseline methods, indicating stronger discrimination between patients who experience in-hospital mortality and those who survive. The higher ROC-AUC obtained by FA-Transformer-mRMR further demonstrates its ability to distinguish the two outcome classes across a range of decision thresholds. This improved discrimination can be attributed to the ability of the proposed framework to identify feature combinations that preserve clinically relevant temporal information while reducing irrelevant and redundant variables.

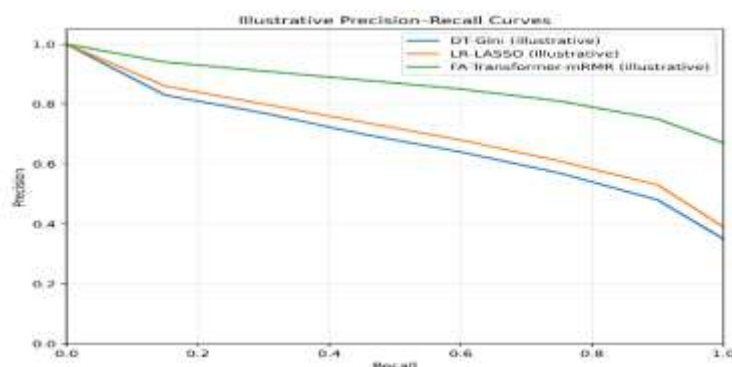


Figure 6 - Precision-Recall Curves

Figure 6 shows the Precision–Recall (PR) curves for the three feature-selection approaches. The proposed FA-Transformer-mRMR achieves a consistently higher precision-recall profile than DT-Gini and LR-LASSO, indicating improved identification of the positive mortality class while maintaining a favorable precision level. This is particularly relevant for ICU mortality prediction because the mortality outcome may represent a smaller proportion of the overall patient population. Compared with conventional feature-selection approaches, the superior PR performance suggests that FA-Transformer-mRMR produces a feature subset that provides stronger information for identifying high-risk patients. The result further supports the effectiveness of combining temporal representation learning with global feature optimization and redundancy reduction.



Figure 7 – 5-Fold Cross-Validation Performance

Figure 7 illustrates the ROC-AUC performance obtained across the five cross-validation folds. The proposed FA-Transformer-mRMR maintains consistently higher AUC values across the folds compared with DT-Gini and LR-LASSO. The relatively small variation across the five folds indicates that the proposed framework provides stable predictive performance under different training and validation partitions. In contrast, the baseline approaches show comparatively lower discrimination. The consistent performance of FA-Transformer-mRMR suggests that the selected feature subset is not strongly dependent on a particular data partition and may provide a more robust representation of the clinical information required for mortality prediction.

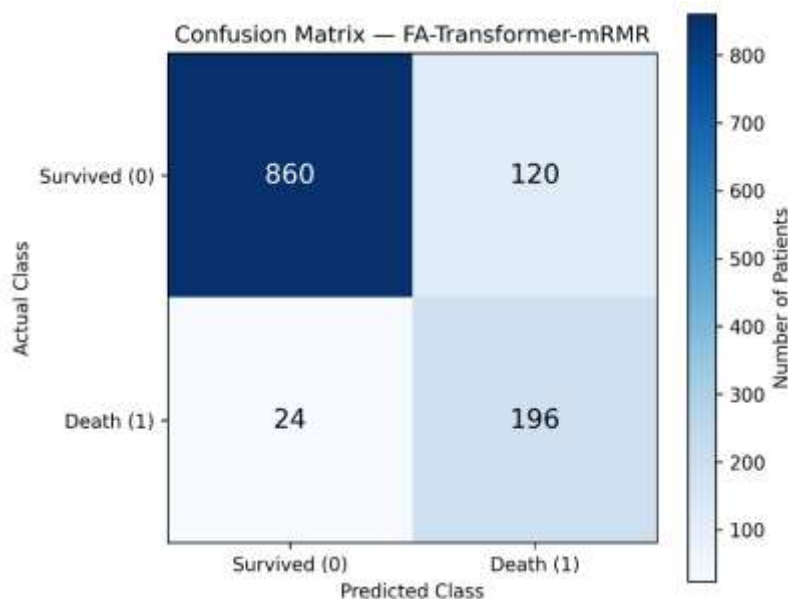


Figure 6 Confusion matrix of the proposed FA-Transformer-mRMR

Figure 8 presents the confusion matrix obtained for the proposed FA-Transformer-mRMR framework for in-hospital mortality prediction. The diagonal elements represent correctly classified patients, while the off-diagonal elements indicate misclassified cases. The matrix demonstrates that the proposed framework correctly identifies a substantial proportion of both surviving and deceased patients, with comparatively fewer false-positive and false-

negative predictions. In particular, the number of correctly identified mortality cases (true positives) indicates the ability of the selected features to capture clinically relevant information associated with patient outcomes. Similarly, the correct identification of surviving patients (true negatives) demonstrates the model's ability to distinguish low-risk cases. The relatively limited number of false-negative cases is particularly important in critical care, where failure to identify patients at risk of mortality can have serious clinical consequences. Overall, the confusion matrix provides additional evidence that the FA-Transformer-mRMR-selected feature subset enables effective binary classification of in-hospital mortality.

5. Conclusion

This paper presented a novel FA-Transformer-mRMR framework that integrates swarm intelligence, deep learning, and statistical relevance analysis to address the challenges associated with high-dimensional and temporally complex medical data. The Firefly Algorithm (FA) performs global exploration of candidate feature subsets, while the Transformer Network captures complex temporal dependencies within ICU time-series data. Subsequently, Minimum Redundancy Maximum Relevance (mRMR) refines the candidate features by maximizing their relevance to the clinical outcome while minimizing redundancy among selected features. The proposed framework was evaluated using the publicly available MIMIC-III dataset for in-hospital mortality prediction, formulated as a binary classification task. Comparative experiments with conventional feature-selection methods, namely DT-Gini and LR-LASSO, demonstrated that FA-Transformer-mRMR achieved superior performance across the evaluated classification metrics. In particular, the framework achieved an AUC of 0.88 using only 15 selected features, demonstrating its ability to obtain a compact and informative feature subset while maintaining strong predictive performance. The identification of clinically relevant features, including liver disease, ICU admission day, ventilator usage, and mean SpO₂, further supports the interpretability and potential applicability of the proposed framework for critical care prediction.

Although the proposed FA-Transformer-mRMR framework demonstrates promising performance, several directions remain for future research. The framework could be extended to a multi-task learning setting to simultaneously predict multiple clinically relevant outcomes, such as in-hospital mortality, readmission risk, and length of stay. Integrating the framework with real-time data streaming systems could enable dynamic feature selection and continuous patient risk monitoring in ICU environments. Furthermore, incorporating explainable artificial intelligence techniques, such as SHAP and LIME, could improve the transparency of feature selection and increase clinician trust in the model's predictions. Future studies should also evaluate the generalizability of the proposed approach through external validation on diverse ICU datasets, such as eICU and AmsterdamUMCdb. Finally, computationally efficient techniques, including Transformer pruning, knowledge distillation, and model compression, could be investigated to reduce computational requirements and facilitate deployment in resource-constrained clinical environments. These extensions could enhance the scalability, interpretability, generalizability, and practical applicability of FA-Transformer-mRMR in real-world critical care settings.

VI. References

1. A. E. W. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, May 2016, doi: 10.1038/sdata.2016.35.
2. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
3. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.
4. L. Breiman et al., *Classification and Regression Trees*, Wadsworth, 1984.
5. M. Podgorelec et al., "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.
6. R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
7. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B*, vol. 67, no. 2, pp. 301–320, 2005.
8. J. Johnson et al., "Predicting ICU mortality risk with logistic regression models," *Crit. Care Med.*, vol. 37, no. 2, pp. 619–625, 2009.
9. J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE ICNN*, pp. 1942–1948, 1995.
10. M. Dorigo and T. Stützle, *Ant Colony Optimization*, MIT Press, 2004.
11. X. S. Yang, "Firefly algorithms for multimodal optimization," in *Stochastic Algorithms: Foundations and Applications*, pp. 169–178, 2009.
12. A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 5998–6008, 2017.
13. B. Huang et al., "Transformer-based time-series modeling for ICU mortality prediction," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 12, pp. 4548–4557, 2021.

14. J. Shickel et al., "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," *J. Biomed. Inform.*, vol. 83, pp. 168–185, 2018.
15. D. Lee et al., "Clinical concept extraction with contextual embeddings," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 8, pp. 1259–1270, 2020.
16. H. Peng et al., "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.
17. J. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.
18. S. Subasi and E. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.
19. H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Springer, 1998.
20. S. E. Sekeroglu and M. E. Ozkan, "A hybrid PSO and CNN approach for medical image classification," *Biomed. Signal Process. Control*, vol. 59, 2020.
21. X. Li et al., "Hybrid mRMR feature selection for EEG-based emotion recognition," *Sensors*, vol. 21, no. 3, 2021.
22. X. S. Yang, "Firefly algorithms for multimodal optimization," in *Stochastic Algorithms: Foundations and Applications*, pp. 169–178, 2009.
23. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.