RESEARCH ARTICLE

A Novel Interpretable Lightweight Ensemble Learning Method for Static and Dynamic Medical and Healthcare Data Classification

Bo Sun[1] and **Hua-Liang Wei**[1,2,3,*]

[1] Department of Automatic Control and Systems Engineering, School of Electrical and Electronic Engineering, The University of Sheffield, Sheffield, United Kingdom
[2] NSIGNEO Institute for in Silico Medicine, The University of Sheffield, Sheffield, United Kingdom
[3] Centre of Machine Intelligence (CMI), The University of Sheffield, Sheffield, United Kingdom

Abstract

In the medical field, efficient and accurate classification of sequential and structured data is crucially important and useful for early diagnosis and treatment. Traditional machine learning models struggle with the complexity and nonlinearity of dynamic datasets, whereas deep learning models, despite their effectiveness, require extensive resources and lack transparency. This paper proposes a novel lightweight ensemble framework integrating a parameterized SoftMax function with a non-parametric Random Forest method through a soft voting mechanism, supported by the Nonlinear AutoRegressive eXogenous (NARX) model and optimized using a forward orthogonal search and selection (FOSS) algorithm for feature selection. This innovative approach enhances the accuracy and robustness of classifiers for both static and dynamic medical datasets, while improving interpretability and computational efficiency. Extensive validation on various medical datasets demonstrates the model's superior performance and adaptability, offering a reliable solution for complex medical data scenarios. It is expected that the results achieved in this study paves the way for future innovations in medical data analysis and the broader application of artificial intelligence in healthcare.

Keywords: healthcare, medical data, static data, dynamic data, classification, soft voting, machine learning.

1 Introduction

In the medical field, efficient and accurate classification of high-dimensional datasets is essential for improving early diagnosis and treatment of diseases. Common clinical and research datasets may be generally divided into two categories: static datasets and dynamic datasets. Static datasets usually include physiological indicators measured at a specific time point, such as blood pressure, blood sugar or cholesterol levels, which are essential for judging the patient's

Academic Editor:
Dharmalingam Muthusamy

Submitted: 02 July 2025
Accepted: 11 August 2025
Published: 16 August 2025

Vol. 2, **No.** 3, 2025.
10.62762/TETAI.2025.713474

***Corresponding author:**
✉ Hua-Liang Wei
w.hualiang@sheffield.ac.uk

Citation
Sun, B., & Wei, H. (2025). A Novel Interpretable Lightweight Ensemble Learning Method for Static and Dynamic Medical and Healthcare Data Classification. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(3), 131–147.

immediate health status and formulating treatment plans. Dynamic datasets, such as electrocardiogram (ECG) and electroencephalogram (EEG), record physiological signals that change over time, providing key information for dynamic monitoring of health status and diagnosis of chronic diseases [1].

Traditional machine learning models perform well in processing and classify static data [2], but they often fail to adapt to the complexity and nonlinear characteristics of dynamic data and struggle to effectively capture the time dependencies and complex patterns hidden in the data. Deep learning techniques have been proven to be effective in processing complex dynamic datasets [3], but they usually require a large amount of data and computing resources. Additionally, the black-box nature of most deep neural network structures makes it less than ideal in some real applications that require fast, interpretable and clear decisions [4]. Therefore, to overcome these drawbacks, this paper proposes an innovative lightweight framework that combines a parameterized SoftMax function and a non-parametric Random Forest model through a soft voting mechanism and operates under the support of the Nonlinear AutoRegressive with eXogenous input (NARX) model, which is a special case of the well-known NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous inputs) representations [5, 6]. To make the generated models transparent, interpretable, parsimonious (compact) and simple (TIPS) , a forward orthogonal search and selection (FOSS) technique is introduced to optimize feature selection. The proposed approach not only improves the accuracy and robustness of the models in handling static and dynamic medical and healthcare dataset classification tasks, but also improves the accuracy and reliability of diagnosis through precise feature selection, which is crucial for high-risk medical and healthcare applications.

This paper will introduce the design and implementation of the ensemble model in detail, evaluate its performance on multiple medical datasets, and explore its potential impact in practical applications. Through comparative analysis with existing technologies, we will show how this method can provide more accurate and interpretable diagnostic results while maintaining efficient data processing. Ultimately, this study aims to provide a feasible and efficient solution for efficient processing of statics or dynamic high-dimensional datasets, paving the way for future medical and healthcare technology innovations.

The contributions of the paper are fourfold:

1. Development of a novel ensemble learning framework based on NARX.
   We developed an innovative ensemble learning framework that integrates nonlinear autoregressive exogenous input (NARX) models with SoftMax and Random Forest models through a soft voting mechanism. This approach not only leverages the strengths of each model in processing data but also significantly enhances the accuracy and robustness of the classifiers when dealing with high-dimensional static and dynamic datasets.

2. Feature information mining.
   Within the proposed ensemble model, the NARX model learns complex features through polynomial combinations, markedly improving the model's ability to detect changes in high-dimensional dynamic data. Moreover, we developed a forward orthogonal search and selection (FOSS) method based on the analysis of variance (ANOVA) technique to optimize feature selection. This method significantly reduces computational complexity and enhances data processing speed by accurately identifying the most impactful features. These enhancements not only boost the model's prediction accuracy with complex time series data but also increase its adaptability and interpretability when handling both dynamic and static datasets.

3. Improvement of model interpretability and adaptability.
   The design of our ensemble model fully considers the high requirements for model interpretability in the medical and healthcare field. By transparently demonstrating the interactions and contributions of each model component, we have substantially improved the model's interpretability, enabling medical and healthcare professionals to better understand and trust the model's decision-making process.

4. Extensive empirical validation.
   We conducted extensive validation on a variety of medical and healthcare datasets, covering simple static datasets to complex dynamic multivariate time series. The test results confirmed the efficiency and reliability of the NARX-based ensemble model in actual medical and healthcare

applications, demonstrated its superiority in processing complex datasets, and provided strong support for future medical and healthcare technology innovation and application.

Through the detailed description of these innovations, we clearly demonstrated the important contributions of the research work at the scientific and practical levels, as well as its role in promoting future data analysis technology.

2 Literature review

2.1 Static and dynamic medical and healthcare datasets

With the advancement of medical technology, medical and healthcare datasets are becoming increasingly important in decision-making, especially in terms of accurate diagnosis and real-time health monitoring. As mentioned earlier, these datasets are usually divided into two types: static and dynamic, each with its own unique application scenarios and analysis requirements. Static datasets usually involve physiological indicators measured once, while dynamic datasets record physiological signals that change over time. Although existing technologies can effectively process these data, the optimized processing methods and challenges faced for specific data types still need to be further explored [7].

In order to have a more systematic understanding of these datasets and their processing technologies, the main characteristics, common physiological signals, challenges and applicable analysis methods related to static and dynamic datasets are summarized in Table 1.

2.2 State-of-the-art in classifying medical and healthcare data

For static data classification tasks, traditional machine learning techniques such as support vector machines (SVM), decision trees, and logistic regression have been widely used, providing disease diagnosis and patient risk assessment based on simple indicators [8]. These methods are favoured due to their simple models and high computational efficiency and are particularly suitable for processing data that do not require complex temporal dependencies. However, they perform poorly when processing dynamic datasets where different features are associated with or dependent on each other temporally, such as electrocardiograms (ECGs) and electroencephalograms (EEGs), which require models to capture long-term dependencies and nonlinear features. Traditional machine learning techniques usually cannot naturally process time series data and require tedious feature engineering to reveal the temporal dependencies in the data, this not only increases the complexity of model building, but may also lead to information loss [9]. Furthermore, these methods generally face challenges in handling nonlinear or high-dimensional datasets, particularly when handling complex interactions between features.

For dynamic datasets, deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and the more recent Transformers are adept at managing complex temporal dependencies typical of time-series data. These models are crucial in predicting patient outcomes efficiently. However, their effectiveness comes at the cost of requiring large amounts of data and substantial

Table 1. Overview of static and dynamic datasets in medical data analysis.

Category	Static Datasets	Dynamic Datasets
Definition	Data measured at specific time points.	Data recorded over a period, capturing changes over time.
Common Signals	Blood pressure, blood sugar, cholesterol levels, etc.	Electrocardiogram (ECG), Electroencephalogram (EEG), etc.
Characteristics	Single-time measurements often include basic physiological parameters.	Continuous recording, displaying trends and variations.
Challenges	Data may lack long-term context, susceptible to measurement noise and errors.	Complex, time-dependent patterns; high data volume increases computational demands.
Techniques	- Logistic Regression - Support Vector Machine (SVM) - Decision Trees - Random Forests - k-Nearest Neighbors (k-NN)	- Recurrent Neural Networks (RNN) - Long Short-Term Memory (LSTM) - Transformer - Nonlinear Autoregressive with Exogenous Input (NARX)

computational resources, which can be limiting in resource-constrained settings. Moreover, the opaque nature of these models poses significant challenges in clinical environments where transparency and explainability are critical for gaining the trust of healthcare professionals and patients. This "black box" aspect often complicates their use in urgent medical scenarios where quick, precise decision-making is necessary, and the reasoning behind every diagnosis or treatment recommendation needs to be clear and justifiable.

When exploring existing data processing methods, the main challenge we face is how to effectively handle various complex datasets, including static and dynamic data, while ensuring the model has high performance and interpretability [10]. This challenge requires us to not only improve the predictive accuracy of the algorithm when designing the model, but also enhance the interpretability of the model to ensure that medical professionals can understand the decision logic of the model. In addition, it is necessary to solve the problems of large-scale data processing and high consumption of computing resources, which is particularly important in resource-constrained medical environments. Therefore, it is particularly important to develop new technical solutions or improve existing technologies to meet these challenges. This includes using hybrid models to make up for their respective shortcomings and developing more efficient training strategies, with the aim of creating an analysis tool that can process data efficiently while being highly transparent and adaptable.

After conducting an in-depth study of the existing technology, we believe that the Nonlinear Autoregressive model with exogenous inputs (NARX) offers a promising solution for addressing complex medical data problems. The NARX model is a dynamic nonlinear system modeling method suitable for data with time dependency and nonlinear characteristics. To accommodate different types of data and application scenarios, various variants of the NARX models have been proposed, such as the polynomial NARX model [6], NARX neural network [11], fuzzy NARX model , and NARMAX model.

The NARX model and Recurrent Neural Networks (RNNs) have similarities in handling time series data but are not the same, nor is NARX a variant of RNNs [12]. NARX (Nonlinear AutoRegressive with eXogenous inputs) is a nonlinear autoregressive model specifically designed to manage dynamic systems with external inputs. It uses lagged input and output values directly for prediction, effectively capturing nonlinear patterns and temporal dependencies in the data. Unlike RNNs, which rely on hidden states to implicitly capture temporal dependencies, NARX depends on explicit historical data and current inputs. This approach enhances the accuracy and flexibility of time series analysis by directly utilizing past inputs and outputs without the need for complex hidden states, thereby effectively managing both dynamic and static datasets.

In standard RNNs (such as simple RNNs, LSTMs, or GRUs), the internal update mechanisms are typically implicit, meaning they do not directly reveal the specific mathematical relationships between inputs and outputs. However, NARX is unique in that its explicit structure directly tracks the relationships between inputs and outputs, providing clear interpretability. This high level of transparency makes the NARX model an excellent tool for gaining insights and providing accurate, flexible decision support in the analysis of complex medical data.

NARX has shown excellent performance in dealing with regression tasks, especially in predicting real-valued outputs such as medical time series data like electrocardiogram (ECG) and electroencephalogram (EEG). However, its ability and application to deal with classification tasks have been relatively less explored, mainly because it was originally designed for prediction using real-valued data recorded sampled from the system of interest. Nevertheless, through in-depth development and innovation of it, we propose applying NARX to classification tasks, which represents a direction with great innovative potential.

In this paper, we propose a lightweight ensemble method that integrates SoftMax and Random Forest with the support of the NARX framework to effectively apply NARX to one-dimensional datasets. Classification tasks. This not only expands the application scope of the NARX model, but also provides an efficient and interpretable tool for processing complex medical datasets, promoting the development of precision medicine and personalized treatment strategies.

3 Methodology

3.1 NARX model

Nonlinear Autoregressive Models with Moving Average and Exogenous Input (NARMAX) methods were initially developed for solving complex nonlinear system identification and modelling problems following systems science and systems engineering principles. A wide range of real-life complex nonlinear systems can be well represented by the NARMAX model. Taking multiple input, single output (MISO) systems as an example, the NARMAX model can be written as:

$$\begin{aligned} y(k) = f\Big(&y(k-1), \ldots, y(k-n_y), \\ &u_1(k-d-1), \ldots, u_r(k-d-1), \\ &\xi(k-1), \ldots, \xi(k-d-n_\xi)\Big) \end{aligned} \tag{1}$$

where f is a linear or nonlinear function; y, u_i $(i = 1, 2, \ldots, r)$, and ξ are the output, the ith input, and noise; $[n_y, n_u, n_\xi]$ are the maximum lags for the system output, input and noise. It is assumed that the noise signal is assumed to be with zero-mean and finite variance. Note that the existing AR, ARX, ARMA, ARMAX, and NARX models can be considered as special cases of NARMAX models, demonstrating the model's applicability to a wide range of systems with nonlinearity and time-varying characteristics.

The NARMAX model is widely used in regression classification problems, especially in time series prediction, such as weather changes, melting of glaciers, prediction of epilepsy, etc., due to its good model structure, the prediction ability and model interpretability. But at present, relatively less efforts have been made to explore the potentials of NARMAX for solving classification problems, which is what we see worthy of innovation.

The NARX model is a subset of the NARMAX model. If we only consider the current noise and ignore its accumulative effect on the system output, that is, we only consider the additive noise at the present time instant, then the NARMAX model (1) reduces to the following NARX model structure:

$$\begin{aligned} y(k) = f\,(&y(k-1), \ldots, y(k-n_y), \\ &u_1(k-d-1), \ldots, u_1(k-d-n_u), \\ &\ldots, u_r(k-d-1), \ldots, u_r(k-d-n_u), \\ &) + \xi(k) \end{aligned} \tag{2}$$

A polynomial NARX model of nonlinear degree ℓ can be expressed in a linear-in-the-parameters form as

$$y(k) = \sum_{m=1}^{M} \theta_m \varphi_m(\varphi(k)) + \xi(k) \tag{3}$$

where $y(k)$ represents the scalar output at time step k, $\varphi_m(\varphi(k))$, with $m = 1, 2, \ldots, M$, are the multivariable polynomial terms that are function of the regressor vector $\varphi(k) = [(y(k-1)), \ldots, y(k-n_y), u(k-1), \ldots]^T$ of past outputs and inputs, θ_m $(m = 1, 2, \ldots, M)$ are the coefficients of the corresponding polynomial terms; and M is the total number of polynomials, the number of polynomial terms, which depends on the nonlinear degree ℓ, and the maximum lags n_y and n_u. M is calculated as follows:

$$M = \binom{n+l}{l} = \frac{(n+l)!}{n! \cdot l!} \tag{4}$$

where $n = n_y + n_u$ in Equation (3) can be rewritten in a vector form as:

$$Y = \phi\theta + \xi \tag{5}$$

3.2 Error Reduction Ratio index and Forward Orthogonal Search and Selection method

Error Reduction Ratio (ERR) is a significant concept used in the field of system identification, particularly in the context of Orthogonal Least Squares (OLS) algorithms. The ERR quantifies the contribution of each potential model term to the reduction of the error variance when the model is included in the model. It is a simple but effective and efficient index for determining the significance of candidate terms. The ERR index of the kth term is calculated as follows.

$$ERR_k = \left(\frac{x_k^T y}{\|x_k\| \|y\|}\right)^2 \tag{6}$$

where y is the observed output vector; x_k is the kth orthogonalized regressor (input vector); $\|x_k\|$ is the Euclidean norm of the kth orthogonalized regressor and $x_k^T y$ is the inner product (dot product) of the kth orthogonalized regressor with the observed output vector.

A Forward Orthogonal Search and Selection (FOSS) algorithm is designed and implemented based on ERR values [13]. Unlike standard forward selection methods, FOSS introduces an element of orthogonality, ensuring that each new feature added to the model is

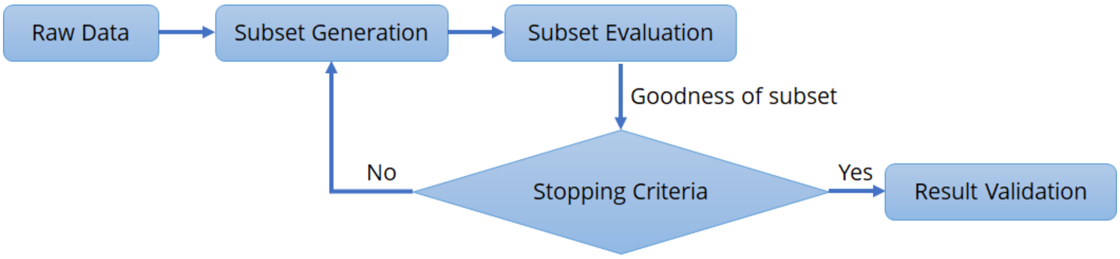


Figure 1. Forward orthogonal search and selection process.

not only significant but also minimally correlated with the features already chosen. This orthogonalization is crucial in preventing multicollinearity, a common pitfall in regression models. Figure 1 shows the diagram of the FOSS algorithm. A detailed description of the algorithm is given in Section 3.3.

3.3 Proposed NARX-based ensemble classifier

3.3.1 NARX-SoftMax

Although the traditional NARX model performs well in capturing the temporal dependencies of dynamic data and handling nonlinear features, it is initially designed for regression problems real-valued outputs and may not be applied to classification tasks directly. To address this issue, we first introduce the SoftMax classifier based on the polynomial NARX model. The SoftMax classifier is a generalization of the logistic regression model [14] that can convert the real-valued output of the NARX model into a discrete probability distribution with a corresponding probability value for each class; this enables the model to be directly applied to multi-class classification tasks. This combination not only leverages the powerful capabilities of NARX in dynamic data analysis, but also makes the model more accurate and reliable in classifying medical conditions through the clear probabilistic output provided by SoftMax.

The NARX model (2) can usually be rearranged to a linear-in-the-parameters form below:

$$\begin{aligned} y(t) = \theta_0 + \sum_{i_1=1}^{n} \theta_{i_1}\phi_{i_1}(t) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \theta_{i_1 i_2}\phi_{i_1}(t)\phi_{i_2}(t) \\ + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_l=1}^{n} \theta_{i_1 i_2 \dots i_l}\phi_{i_1}(t)\phi_{i_2}(t)\dots\phi_{i_l}(t) \\ + \xi(t) \end{aligned} \tag{7}$$

where the parameter ℓ is called the degree of nonlinearity which controls the complexity of the model, where the power of the cross-product terms is not higher than ℓ. The total number of potential terms in the model is given by (4).

For the SoftMax regression, the model adapts to deal with multi-classification tasks by defining an output for each class j as follows:

$$y_j(t) = \sum_{m=1}^{M} \theta_{jm}\phi_m(\varphi(t)) + \xi(t) \quad \text{for} \quad j = 1, \dots, K \tag{8}$$

$$p_j = \frac{e^{y_j}}{\sum_{k=1}^{K} e^{y_k}} \tag{9}$$

Note that for (8), each class j has its own weight vector θ_j. K denotes the total number of classes. Figure 2 shows the workflow of the NARX-SoftMax model.

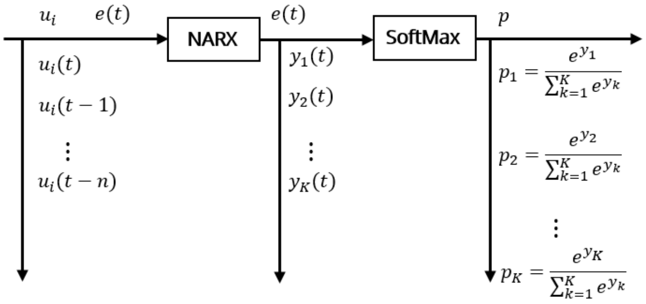

u_i
$e(t)$
NARX
$e(t)$
SoftMax
p
$u_i(t)$
$u_i(t-1)$
$\vdots$
$u_i(t-n)$
$y_1(t)$
$y_2(t)$
$\vdots$
$y_K(t)$
$p_1 = \frac{e^{y_1}}{\sum_{k=1}^{K} e^{y_k}}$
$p_2 = \frac{e^{y_2}}{\sum_{k=1}^{K} e^{y_k}}$
$\vdots$
$p_K = \frac{e^{y_K}}{\sum_{k=1}^{K} e^{y_k}}$

Figure 2. The workflow of the NARX-SoftMax model. The input layer consists of current and past input values $u_i(t), u_i(t-1), \dots, u_i(t-n)$ along with an error term $\xi(t)$. These inputs are processed by the NARX model to produce an intermediate output $y_j(t)$, which is converted into probabilities p_j for each class j by the SoftMax classifier.

3.3.2 Feature selection

In traditional NARX models, the Error Reduction Ratio (ERR) metric is typically employed for feature selection in regression tasks. However, ERR's applicability is limited when it comes to classification tasks, particularly in the context of probabilistic models like SoftMax. To address this limitation, we turn to Analysis of Variance (ANOVA), a statistical technique that excels in identifying significant features for classification purposes. Specifically, in this setting, F-statistic for each feature used to replace ERR as the primary tool for assessing the relevance of features. It provides a robust framework for testing the statistical significance of each feature, determining how different levels of a feature affect class probabilities. This approach is particularly effective in scenarios where the relationship between features and class outcomes is not linear or straightforward. F-Statistic is defined as follow:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}} \tag{10}$$

Between-Group Variance represents the variance of the group means from the overall mean. Within-Group Variance reflects the average of the variances within each group. The F-statistic thus provides a measure of the ratio of the explained variance between groups to the unexplained variance within groups. A higher F-statistic indicates a greater disparity between group means, suggesting that the variable under consideration has a significant effect on the outcome variable. In hypothesis testing, this measure is used to reject or fail to reject the null hypothesis that the group means are equal, with a lower p-value indicating a higher likelihood of the means being significantly different.

3.3.3 *Stopping criteria for forward feature selection*

In order to improve the performance of the classification model and avoid overfitting, we developed the Penalized Error-to-Signal Ratio (PESR) as the stopping condition in the feature selection process. PESR combines the prediction error and complexity of the model and controls the complexity of the model by introducing a penalty term. In this paper, we use cross-entropy loss as a measure of model prediction error and gradually add features through forward selection to optimize the performance of the classification model. When the number of important features selected is k, the PESR is defined as follows:

$$PESR_k = \left(\frac{N}{N - \lambda k}\right)^2 \times Loss_k \tag{11}$$

The cross-entropy loss is defined as below:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(P_{ij}) \tag{12}$$

where N is the number of samples in the test set, k is the current number of features in the model, λ is a parameter controlling the penalty term (which is usually chosen to be $\lambda \geq 1$), and $Loss_k$ is the cross-entropy loss at the k-th iteration.

When the inclusion of an additional new feature cannot help reducing the PESR value, the iteration terminates, meaning that the current feature combination has reached the optimal balance point, as adding more features will only increase complexity without significantly improving the model prediction performance.

PESR can effectively control the model complexity; it is particularly effective in high-dimensional datasets or when dealing with many features. Using cross-entropy loss as the error measure provides a more comprehensive evaluation of model performance compared to accuracy.

The Forward Orthogonal Search and Selection (FOSS) algorithm [6, 13], which is used to implement the feature selection procedure for the NARX-SoftMax model described in Sections 3.3.1-3.3.3, is depicted in Algorithm 1.

In the algorithm, the initial step involves computing the weight w_i for each regressor vector ϕ_i in the dictionary, which helps identify the candidate model terms with the highest contribution. The algorithm then selects the term with the highest weight and initializes the NARX-SoftMax model using this term, followed by training and performance evaluation.

Subsequently, the algorithm iteratively uses the modified Gram-Schmidt method to orthogonally select the next candidate term that contributes the most to the model, while removing terms with minimal contribution. Each selected term is evaluated through ANOVA. This process continues until the model's accuracy no longer improves, indicating that additional terms would no longer enhance the model. This procedure follows a greedy approach, selecting the best candidate term at each iteration.

Finally, the algorithm outputs the selected important feature subset, which represents the terms that contribute most significantly to the NARX-SoftMax model.

3.3.4 *Enhancing model robustness and reliability with Random Forest ensembles*

To further improve the NARX-SoftMax model performance, a soft voting mechanism [15] combined with Random Forest (RF) is proposed in this paper to enhance the robustness and reliability of the classifier and the interpretability of the entire model system.

Algorithm 1: FOSS for NARX-SoftMax model

```
Input: Dictionary of regressor vectors
       $D = \{\phi_1, \phi_2, \ldots, \phi_M\}$, output signal $y$,
       maximum number of terms $t_{max}$
Output: NARX-SoftMax model with significant
        terms selected from $D$
for all $\phi_i \in D$ do
    Define $w_i = \|\phi_i\|$ ;
end
Compute $F(\phi_i, y)$ ;
Define $j = \arg\max \|w_i\|$ ;
Define $q_1 = \phi_j$ ;
Define $p = \phi_j$ ;
Train a NARX-SoftMax model using $y$ and $p_1$ ;
Remove $\phi_j$ from $D$ ;
for $s = 2$ to $t_{max}$ do
    for all $\phi_i \in D$ do
        Orthogonalize $\phi_i$ with respect to $[q_1, \ldots, q_s]$
         to obtain $w_i$ ;
        if $w_i^T w_i < 10^{-10}$ then
            Remove $\phi_i$ from $D$ ;
            Go to next iteration ;
        end
    end
    Compute $F(\phi_i, y)$ ;
    Find $j = \arg\max_{i=1}^{s-1} F(\phi_i, y)$ ;
    Define $q_s = \phi_j$ ;
    Define $p_s = \phi_j$ ;
    Train a NARX-SoftMax model using $y$ and
     $p_1, \ldots, p_s$ ;
    Compute Accuracy ;
    if the PESR value no longer reduces then
        Delete $\phi_j$ from $D$ ;
        Go to next iteration ;
    end
    Record Accuracy and Remove $\phi_j$ from $D$ ;
end
return Matrix of terms selected $P = [p_1, p_2, \ldots, p_n]$
```

Random Forest can handle nonlinear relationships and complex interactions between features well, provides additional error correction capabilities, thereby improving the robustness of the model and adapting to complex classification tasks [16]. As shown in Figure 3, a random forest is made up of several decision trees, each of which predicts the class for a new sample. In this example, three trees (Result 1, Result 2, and Result 3) provide their own predictions. The final result is determined by majority voting, where the most common prediction among the trees is chosen. This method helps improve accuracy by combining the strengths of multiple trees to make the final prediction.

Soft voting is an advanced aggregation technique used in our ensemble model to combine the predictive strengths of the SoftMax and RF modules. Each module generates a number of probabilities for each class after processing the input data. Soft voting takes these probabilities into account rather than just considering the most voted class. This ensemble strategy improves the overall predictive accuracy because it integrates the confidence levels of each classifier's output into the decision-making process. By averaging the probabilities, the model reduces the variance and potential biases that might occur if only a single classifier's output was considered. This is particularly beneficial in medical data analysis, where the stakes are high, and decisions need to be both accurate and reliable. Furthermore, soft voting inherently provides a layer of error correction, as it tends to cancel out any outlier predictions made by individual models in the ensemble. For example, if one model erroneously predicts a rare class due to some noise in the data, but the other models consistently predict a more probable class, the final averaged probabilities will likely favour the correct class. Figure 4 shows the overall structure of the proposed ensemble model NARMAX-SoftMax-RF (shortly NARX-SR), where 'S' represents SoftMax, and 'R' represents Random Forest.

This flowchart illustrates the overall structure of the NARMAX-SoftMax-RF (NARX-SR) model. First, the input data is processed through the NARX model to generate polynomial features, followed by feature selection using FOSS combined with ANOVA. The filtered feature set is then used to train the SoftMax and Random Forest (RF) classifiers. The output probabilities from these classifiers are combined using a soft voting mechanism to produce the final prediction. The process checks if the PESR value is reduced to determine whether further feature selection optimization is required. The process then checks if the PESR value is reduced. If yes, the model proceeds to predict on the test set and evaluates the results. If not, further optimization is done before making predictions.

4 Case study

To evaluate the performance of the proposed NARX-based classifier, we conducted a series of case studies relating to classification tasks focused on

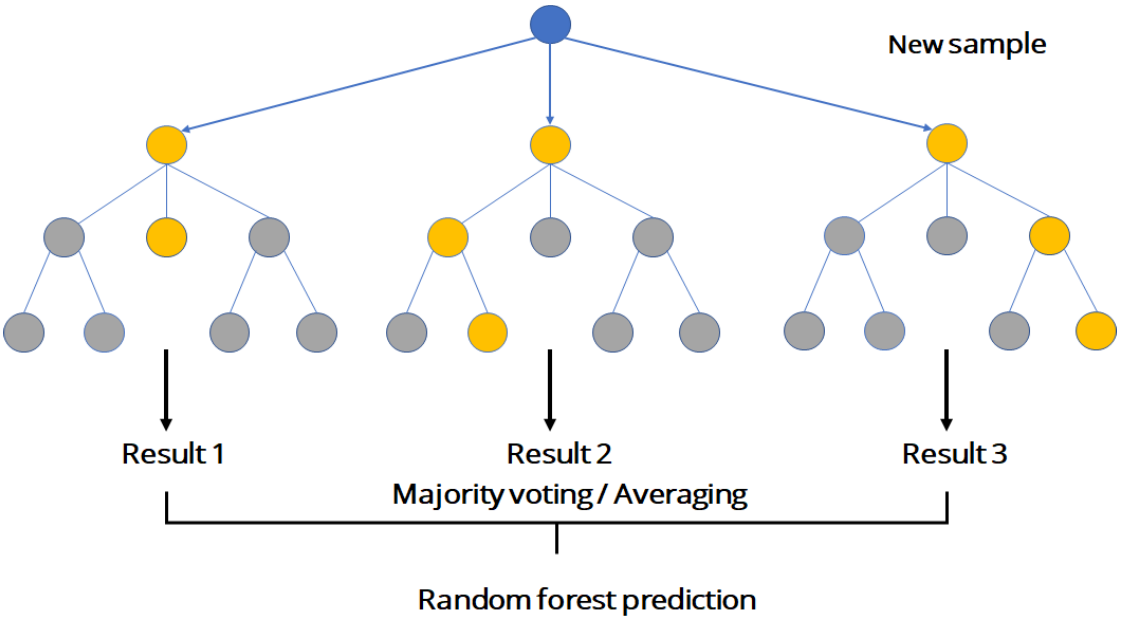


Figure 3. Prediction process in a random forest model. The diagram illustrates how the soft voting mechanism integrates the probability outputs from multiple classifiers to make a final decision, enhancing accuracy and robustness.

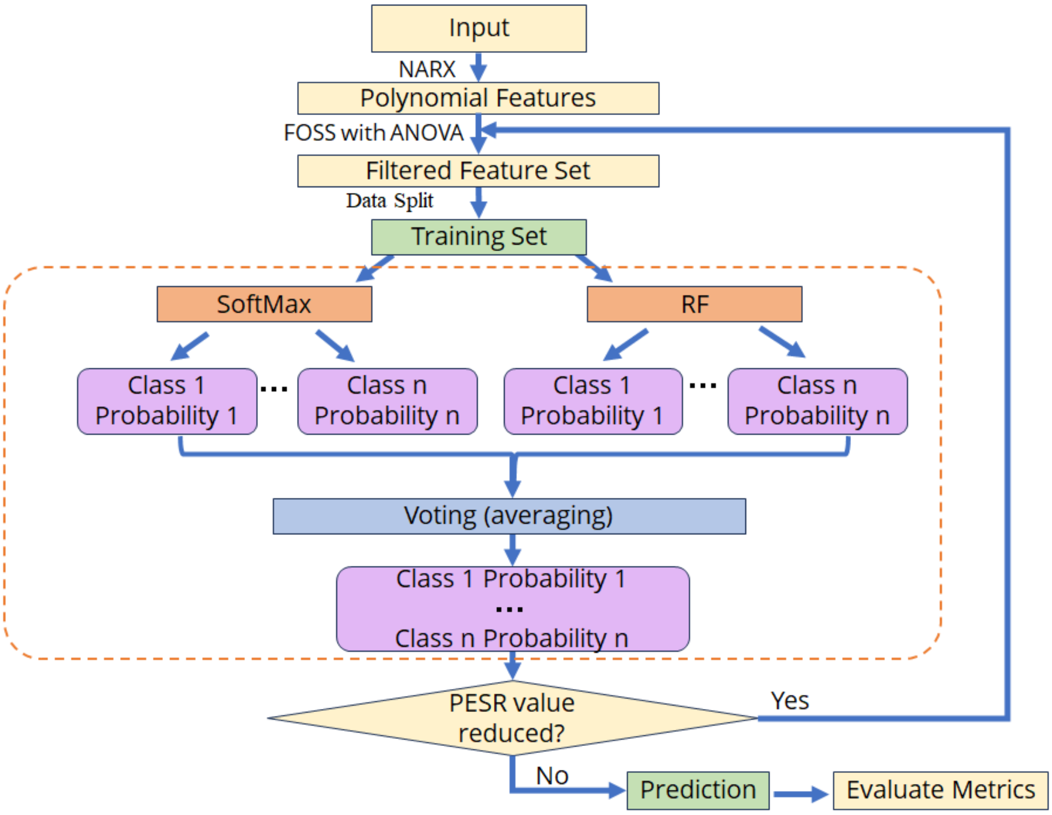


Figure 4. The overall structure of the proposed 'NARX-SR' ensemble framework.

both static and dynamic datasets. The aim was to assess the classifier's prediction accuracy and feature selection capability across different scenarios. We started with a traditional static classification task as an initial test. Following this, we utilized a time series dataset for binary classification to examine the classifier's dynamic capabilities. Lastly, we applied our method to a multi-class classification problem involving a dynamic dataset of Alzheimer's Disease (AD). This progression allowed us to test the method's adaptability and performance across varying classification challenges.

We evaluated the classification performance using four commonly used key metrics: accuracy, precision, recall, and F1 score [17]. Accuracy is an index showing the overall proportion of correct predictions, while precision and recall provide insights into the model ability to correctly identify specific categories—precision focusing on the accuracy of positive predictions and recall on their completeness. The F1 score combines precision and recall, offering a balanced perspective. Additionally, we detailed the feature subset selected by our method, highlighting its impact on enhancing the model performance.

4.1 Case 1

The dataset employed in the first case study is designed to facilitate the estimation of obesity levels among individuals from Mexico, Peru, and Colombia, utilizing variables related to their dietary habits and physical condition [18]. We selected this dataset as our initial case study because it represents a typical static multiclass dataset within the medical field. Additionally, its medium size makes it ideally suited for both conducting an initial analysis of our developed model and comparing it with other classic models. The dataset comprises 2,111 instances and 17 attributes, aimed at studying obesity level classification. The 'Obesity' variable classifies individuals into seven distinct obesity levels: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Approximately 77% of the data was synthesized using the Weka software with the SMOTE filter, while around 23% of the data was directly obtained from individuals via a web-based platform.

In this case, we performed preprocessing on the selected dataset, including filling missing values and dividing 70% of the data into a training set which we employed 10-fold cross-validation on it and the remaining part as a test set. In particular, considering that the dataset does not contain lag terms, we manually set the nonlinearity of the model to 2 based on previous experience to adapt to the characteristics of the data. This process involves a search space of 171 features designed to optimize model performance.

Table 2 shows the performance comparison of various classification methods on the test set. The NARX-SR method achieved the highest scores, with an accuracy of 0.9779, precision of 0.9792, recall of 0.9773, and an F1 value of 0.9746. This superior performance is due to the Soft Voting method based on the NARX model, which effectively combines multiple classifiers to reduce bias and variance. Other models, such as NARX-SoftMax, NARX-RF, and RF, did not perform as well as the NARX-SR (NARMAX-SoftMax-RF) ensemble. While CNN and LSTM generally excel with larger datasets, they underperformed on this small dataset. SVM with L1 regularization, DT, and KNN showed reasonable results but did not match the NARX-SR model.

Table 2. Results of case1(on test set). NARX-SR: NARMAX + SoftMax + Random Forest.

Methods	Accuracy	Precision	Recall	F1
NARX-SoftMax-RF	**0.9779**	**0.9792**	**0.9773**	**0.9746**
NARX-SoftMax	0.9605	0.9244	0.9235	0.9237
NARX-RF	0.9557	0.9487	0.9484	0.9483
SVM	0.9510	0.9527	0.9485	0.9490
KNN	0.7930	0.7855	0.7781	0.7763
RF	0.9494	0.9487	0.9484	0.9483
DT	0.8262	0.8219	0.8357	0.8212
CNN	0.9052	0.9017	0.9006	0.9005
LSTM	0.9368	0.9365	0.9356	0.9346

Our developed model not only exhibits excellent predictive performance but also excels in interpretability and identifying and ranking important features and their interactions within the dataset. Figure 5 shows the accuracy improvement of our model as important features are added, achieving a peak accuracy of 97.79% with a minimal but crucial feature set. This demonstrates the model's high learning ability and efficiency. Figure 6 displays the confusion matrix at the model's best performance.

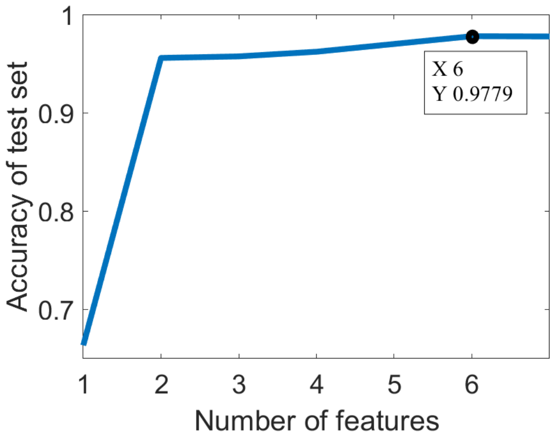
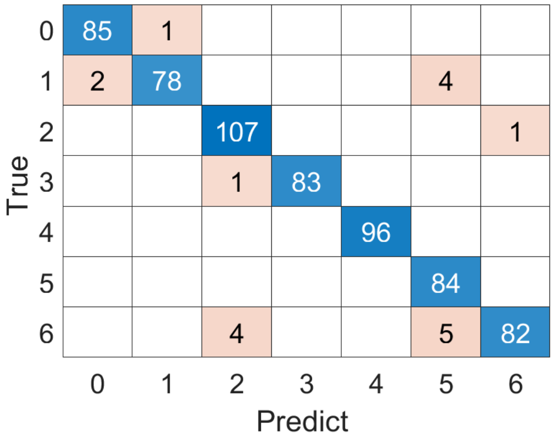


Figure 5. Accuracy plot of NARX-SR (test set).

Table 3 presents the selected important feature subset, including both original features and nonlinear combinations (Weight, Height×Height, Gender×Weight, Gender×Height, Age×family_history_with_overweight, FAVC×NCP). This proves that our method effectively captures complex data patterns, provides interpretable



	0	1	2	3	4	5	6
0	85	1					
1	2	78				4	
2			107				1
3			1	83			
4					96		
5						84	
6			4			5	82

Figure 6. Confusion matrix of the NARX-SR predictions (test set).

Table 3. The selected features for Case 1. From top to bottom, the importance of features decreases.

Features' ID
Weight
Height × Height
Gender × Weight
Gender × Height
Age × family_history_with_overweight
FAVC × NCP

information for model predictions and enhancing model performance.

Figure 7 Comparisons of different methods for feature selection. We used Random Forest and l1-regularized SVM as the baselines to evaluate the performance of the proposed NARX-SoftMax and NARX-RF. The NARX-SR (NARMAX-SoftMax-RF) model consistently shows the highest accuracy, achieving optimal performance with just six important features, demonstrating its robustness and effectiveness in feature selection.

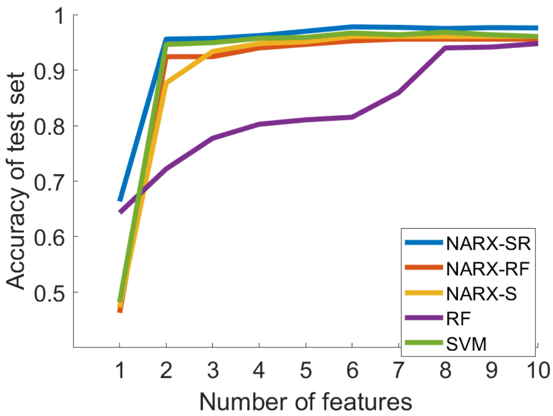


Figure 7. A comparison of different methods in feature selection (test set).

Table 4. Description of eye state dataset.

Attribute	Details
Dataset Name	Eye State dataset
Source	UCI Machine Learning Repository
Classification Type	Binary (Dynamic classification)
Measurement Device	Emotive EEG Neuroheadset
EEG Channel Names	AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8
Measurement Duration	117 seconds
Eye State Detection	Detected via camera during EEG measurement, added manually after video analysis
States Encoding	'1' for eye closed, '0' for eye open
Data Ordering	Chronological

4.2 Case 2

This is a case concerned with a binary classification task based on dynamic data. For the second case study, we selected a binary classification time series dataset - the Eye State dataset from the UCI Machine Learning Repository [19] as our case study of binary dynamic classification. Details are in the Table 4:

For this EEG dataset, past time information is added to each sample point; here, the time delay is set to 5 and the nonlinear degree is set to 2. Thus, we can scramble all samples for dynamic analysis without losing time information. There are a total of 3655 candidates in the search space, and similar to the breast cancer example, we select 70% as our training set which we employed 10-fold cross-validation on it and 30% as our test set. In they employed the PCA to reduce the dimensionality of the dataset, which is not what we desire. Our objective is to identify the association between the original EEG channel data and the results and to provide an interpretation based on this information. As in Case 1, we used the same models as the baselines for comparison purposes. Note that the same autoregressive inputs are used in all the compared methods.

Unlike for the above Case 1 where data are static, for the time series binary classification task here, the performance of different methods varies significantly. As Table 5 shows, the NARMAX-SoftMax-RF method remains highly stable and achieves the highest performance due to its effective integration of nonlinear time series features and complex temporal dynamics. NARX-RF and RF also perform well but fall short of the NARX-SR method, which uses a soft voting ensemble approach. LSTM performs well due to its strong ability to capture temporal dependencies but has drawbacks such as long training time, high computational resources, being a black-box model, and

Table 5. The results of the test data given by different methods on the test data for Case 2.

Methods	Accuracy	Precision	Recall	F1
NARX-SoftMax-RF	**0.9034**	**0.9061**	**0.9018**	**0.9035**
NARX-SoftMax	0.6693	0.6539	0.6696	0.6536
NARX-RF	0.8823	0.8827	0.8901	0.8854
SVM	0.5895	0.5439	0.6173	0.4822
KNN	0.7766	0.7663	0.7787	0.7693
RF	0.8727	0.8684	0.8743	0.8705
DT	0.7183	0.7109	0.7147	0.7122
CNN	0.6233	0.6274	0.6266	0.6231
LSTM	0.8117	0.8190	0.8006	0.8048

complex hyperparameter tuning. Methods like SVM with L1 regularization, CNN, and NARX-SoftMax perform poorly due to their limitations in handling complex time series data. Overall, the NARX-SR method is recommended for the time series binary classification tasks due to its superior performance. Case 2, the proposed NARX-SR model demonstrates a strong feature selection capability. Table 6 briefly shows the critical feature interactions captured by the model. With these interactions, the model's accuracy improved to a maximum of 0.9034. Figure 8 illustrates this trend, showing a steady increase in accuracy with the number of features, peaking at the 15th feature. The confusion matrix displayed in Figure 9 details the distribution of model prediction results.

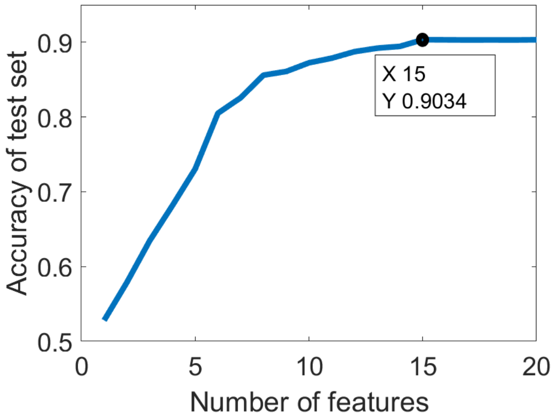


Figure 8. Accuracy plot of NARX-SR (test set).

In Table 6, F7(t) × F7(t-1) represents the multiplication of the current signal at the F7 electrode with the signal at the same electrode from the previous time step (t-1), capturing temporal dependencies between consecutive time points. Similarly, T8(t-1) × FC6(t-1) indicates the interaction between signals from the T8 and FC6 electrodes at the previous time point, and F3(t-1) × T7(t-1) reflects the interaction between F3 and T7 from the previous time step. All these interactions help model temporal and spatial relationships in EEG data to enhance classification. To verify our approach, we compared it with other feature selection methods. As shown in Figure 10, NARX-SR (NARMAX-SoftMax-RF) outperforms all other methods, including NARX with L1 regularization and NARX-SoftMax, which performed relatively poorly. Random Forest (RF) and our proposed NARX-RF also performed well, but NARX-SR showed the best performance. NARX-SR excels in handling complex time series data by combining the strengths of parametric models (like NARX) and non-parametric models (like RF). Parametric models may struggle with complex patterns, while non-parametric models handle non-linear relationships and high-dimensional data effectively. By integrating both, NARX-SR achieves optimal performance.

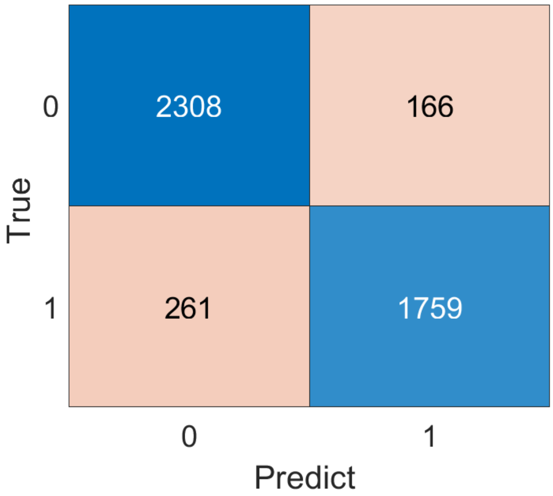


Figure 9. Confusion matrix of the NARX-SR predictions (test set).

Table 6. Selected important feature subset (Channels and Sampling Points) for Case 2. From top to bottom, the importance of features decreases.

Features' ID
F7(t) × F7(t-1)
T8(t-1) × FC6(t-1)
...
...
...
F3(t-1) × T7(t-1)

4.3 Case 3

This case is concerned with a dynamic multi-class classification problem, where the time series data contains dynamic relationships determined by the underlying nonlinear patterns and sequence-dependent properties. For such a multi-class classification problem, we focus on the model accuracy and its ability to utilize sequential information and

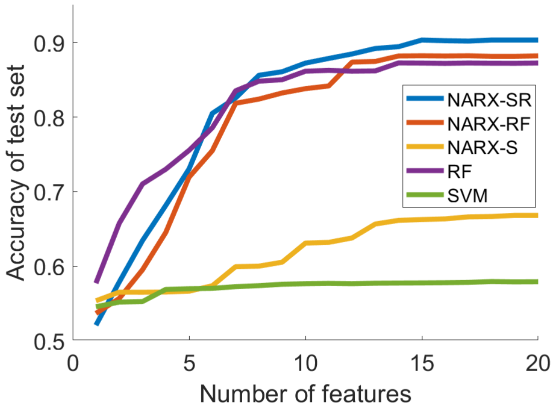


Figure 10. A comparison of different methods in feature selection (test set).

Table 7. Description of eye state dataset.

Attribute	Details
Source	Florida State University
Scalp Locations	19 locations (Fp1, Fp2, Fz, F3, F4, F7, F8, Cz, C3, C4, T3, T4, Pz, P3, P4, T5, T6, O1, O2)
System	International 10-20 system
Equipment	Biologic Systems Brain Atlas III Plus workstation
Cerebral Lobes	F: Frontal, C: Central, P: Parietal, O: Occipital, T: Temporal
Groups	A and C: Eyes open, B and D: Eyes closed
Participants	Groups A and B: 24 healthy elderly (Avg age 72, range 61-83), Groups C and D: 24 probable AD patients (Avg age 69, range 53-85)
Diagnosis Criteria	NINCDS, ADRDA, DSM-III-R
EEG Segment Duration	8 seconds
Frequency Range	1-30 Hz
Sampling Frequency	128 Hz
Data Integrity	Free from eye motion and blinking, myogenic artifacts
Monitoring	EEG technician present during recordings

capture important characteristics. To this end, we considered the EEG database provided by Florida State University [20]; This dataset includes recordings from 19 scalp locations using the International 10-20 system, with participants grouped into healthy controls and probable AD patients. The details of this EEG dataset are summarized in Table 7.

This dataset is particularly valuable for studying Alzheimer's Disease due to its comprehensive coverage of both healthy and AD-affected participants, as well as its detailed EEG recordings under controlled conditions, which makes it ideal for time-series analysis and model validation. EEG data from 8 elderly participants per group were used for training, with the rest as the test set. The model classifies each sampling point, and the diagnosis is made based on a majority voting scheme. The basic modelling experimental settings are as follows: the maximum input and output time lags were set to five, and the degree of the nonlinearity of the NARX model was chosen to be two. The prediction results with different methods are shown in Table 8.

Table 8. Results of case 3 (test set).

Methods	Accuracy	Precision	Recall	F1
NARX-SoftMax-RF	**0.6966**	**0.6894**	**0.6920**	**0.6892**
NARX-SoftMax	0.6143	0.6137	0.6233	0.6148
NARX-RF	0.6793	0.6790	0.6773	0.6765
SVM	0.4782	0.5540	0.4794	0.4768
KNN	0.5223	0.6062	0.5231	0.5129
RF	0.5517	0.5416	0.5458	0.5422
DT	0.4476	0.5045	0.4464	0.4359
CNN	0.6206	0.6189	0.6361	0.6203
LSTM	0.5967	0.5966	0.5980	0.5964

As shown in Table 8, the NARX-SR model achieved the highest performance across all metrics with an accuracy of 0.6966, precision of 0.6894, recall of 0.6920, and an F1 score of 0.6892, making it the most effective model for multi-class time series classification. Other models like NARX-RF, NARX-SoftMax, CNN, and LSTM also performed well, But slightly worse than NARX-SR. Conversely, models such as SVM with L1 regularization and DT showed lower performance across all metrics, indicating they are less suited for this specific classification task. SVM-l1 may struggle with complex decision boundaries, while DT may not effectively capture temporal dependencies in the data.

In Case 3, our model continues to demonstrate robust data exploration and feature detection ability. Figure 11 shows the accuracy plot of NARX-SR, achieving an accuracy of 0.6966 with 43 features. This indicates that the model effectively utilizes the selected feature subset to enhance performance. Figure 12 presents the confusion matrix for NARX-SR predictions, illustrating the detailed classification results across multiple classes. The matrix shows that the model performs well in distinguishing between different classes, with a significant number of true positives in each category. Table 9 briefly lists the important feature combinations selected for Case 3. These interactions reveal the underlying patterns and relationships within the data, further validating the model's ability to handle complex multi-class classification tasks.

In Table 9, $O1(t) \times O1(t)$ indicates the squared value

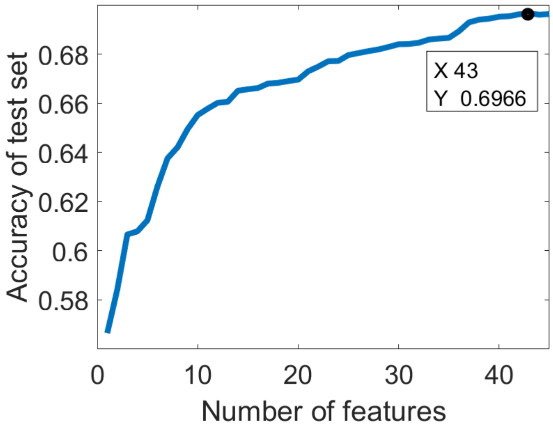


Figure 11. The Accuracy plot of NARX-SR (test set).

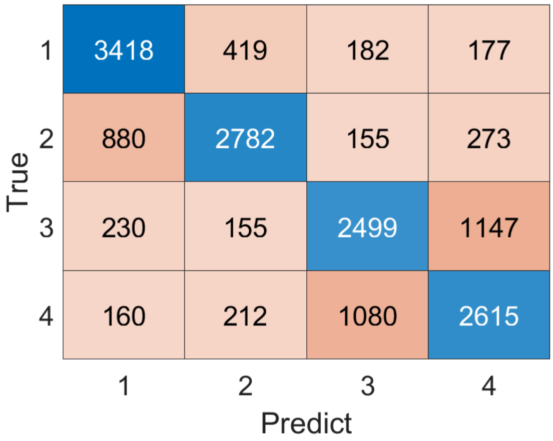


Figure 12. Confusion matrix of the NARX-SR's predictions (test set).

of the signal at the O1 electrode, and O1(t-3) × O1(t-3) refers to the squared value of the O1 signal from three-time steps earlier. Similarly, T3(t-2) × T4(t-2) captures the interaction between signals from the T3 and T4 electrodes, both delayed by two-time steps. These interactions are crucial for capturing both temporal and spatial relationships in the EEG data.

For multi-class classification of time series datasets, extracting useful information efficiently is crucial. As shown in Figure 13, NARX-SR (NARMAX-SoftMax-RF) outperforms NARX-RF, achieving optimal performance with 40 features, while NARX-SoftMax reaches only 69% accuracy. RF and SVM with L1 regularization perform poorly due to their inability to capture complex temporal dependencies and nonlinear features. RF handles nonlinear relationships well but falls short in temporal dependencies, and SVM-L1 struggles with complex decision boundaries.

Table 9. Selected important feature subset (Channels and Sampling Points) of case 3 ((From top to bottom, the importance of features decreases).

Features' ID
O1(t) × O1(t)
O1(t-3) × O1(t-3)
...
...
...
T3(t-2) × T4(t-2)

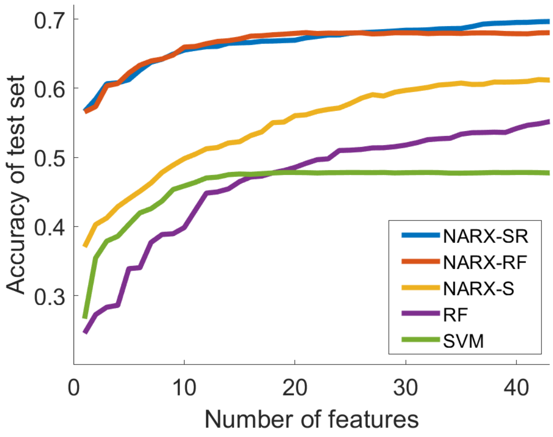


Figure 13. Comparison of different methods in feature selection (test set).

4.4 Computational Complexity and Experimental Fairness

All comparative experiments in this study were implemented and executed by the authors to ensure fairness and reproducibility. For each baseline method, the same training–testing splits, preprocessing procedures, and parameter tuning strategies were applied as those used for the proposed model, ensuring that performance differences arise solely from the modelling approaches.

The complete pipelines — including feature selection using the FOSS algorithm, feature generation via the NARX model, 10-fold cross-validation, and evaluations across all baseline models — were executed on the same three datasets in this study. On a workstation equipped with an Intel i7 CPU, 32 GB RAM, and an RTX 4070 Ti GPU, the total running time for the simpler static datasets was at the second level, while the largest and most complex EEG dataset was completed at the minute level, with all runs taking no more than 10 minutes. These results demonstrate that the proposed approach is computationally efficient and suitable for practical medical data analysis. For substantially larger datasets, the running time may increase, but the method remains scalable due to its modular and parallelizable design.

4.5 Discussions

4.5.1 The advantages of the proposed methods

Three different types of case studies were conducted to evaluate the performance of the proposed methods, by comparing them against several common classifiers widely applied within the medical field. The findings demonstrate that our models showcased superior performance and a degree of interpretability across all cases. In essence, these models are not only adept at handling static classification tasks efficiently but also excel in managing multi-class classification issues in medical data characterized by complex temporal dynamics, such as EEGs or ECGs. This innovative framework is designed to learn nonlinear relationships from data and the framework enhances feature understandability and model interpretability by computing contributions and selecting features after polynomial combinations through forward orthogonal selection.

Furthermore, by integrating the SoftMax classifier and Random Forest using the soft voting method, we can leverage the strengths of both methods. Such an approach provides several benefits in the context of medical dataset, which can be briefly summarised as follows.

Combining Strengths of SoftMax and Random Forest. The SoftMax classifier is excellent for handling linear relationships and providing probabilistic interpretations of class membership. On the other hand, Random Forest is strong in capturing complex, non-linear relationships and handling high-dimensional data with robustness to overfitting. Soft voting works by averaging the predicted probabilities from both the SoftMax and Random Forest models. This method enhances the classification performance by taking advantage of the strengths of each model. In medical data analysis, this leads to better prediction accuracy, which is critical for reliable diagnostic and prognostic outcomes.

Feature Selection and Ranking. Using forward feature selection ensures that the most significant features are included in the modelling process in a step-forward manner. This enhances both the performance and interpretability of the model. In the medical domain, understanding the influence of each feature is crucial, and this method helps in identifying the key variables that impact patient outcomes.

Transparency and Interpretability. The explicit structure of the NARX model allows for clear tracking of the relationships between inputs and outputs. When combined with the interpretability of Random Forest feature importance and the probabilistic outputs of SoftMax, the resulting model is both transparent and understandable. This transparency is vital in medical data analysis, where model interpretability can directly influence clinical decision-making.

Robustness to Data Variability. The integration approach ensures that both static and dynamic aspects of the data are effectively analysed. The polynomial NARX model helps in capturing nonlinear temporal dependencies within the data, while the Random Forest and SoftMax classifiers address various static features. This combination makes the model robust to the variability and complexity inherent in medical datasets.

4.5.2 Limitation and Challenges

Despite the significant progress and strategic advantages of using NARX-based models in healthcare analytics for various classification tasks, challenges remain that require further investigation and development, for example:

Feature Selection and Feature Engineering. A major challenge in healthcare analytics involves the scalability and computational efficiency of models as they are applied to increasingly large and complex patient datasets. Traditional ANOVA-based feature selection methods are effective for medium-sized datasets, may struggle with processing speed and scalability as the volume and dimensionality of data grow. This limitation is critical in clinical settings where rapid processing and timely model updates are essential. Moreover, the reliance on ANOVA's statistical assumptions may not always be valid in medical datasets, because it assumes data follows a normal distribution and equal variances across groups, which may not hold true for all datasets, limiting its applicability. Potentially impacting the model's effectiveness and its applicability across different types of patient data. This highlights the need for more adaptable feature selection techniques that can handle the non-linear and heteroscedastic nature of medical data.

Trustworthy in Decision Making. The benefits of model interpretability and transparency are crucial in a medical context. However, as models and feature selection processes become more complex, creating intuitive visualizations and explanations to demystify model decisions for non-experts is increasingly vital. This is essential for encouraging the adoption and trust in machine learning solutions within healthcare settings.

Addressing these challenges requires a multi-faceted approach, including the development of more scalable and adaptive feature selection methods, exploration of hybrid models that combine the strengths of parametric and non-parametric approaches, and advances in interpretability tools to bridge the gap between complex model outputs and user comprehension. Such efforts are crucial for maximizing the potential of NARX based models in healthcare, leading to more effective, efficient, and user-friendly machine learning applications in patient care and medical research.

5 Conclusion

In this paper, we developed a lightweight ensemble classification framework based on the NARX model, integrating a parameterized SoftMax regression and a non-parametric Random Forest using a soft voting mechanism. This approach effectively handles both static data (e.g., patient demographics and EHR) and dynamic time series data (e.g., EEG) in medical applications. Combined with forward orthogonal search and selection (FOSS) algorithm for feature selection, our method has ability to select the most relevant features and the nonlinear combination relationship of deep mining features, reducing computational complexity while maintaining high classification accuracy.

Our method demonstrated robust predictive performance, efficient feature selection, and improved interpretability across the static and dynamic datasets used in the three case studies. The ensemble model showed strong adaptability and efficiency, enhancing classification accuracy while providing richer interpretive insights into both static datasets, such as patient records, and dynamic datasets, like time-series signals. These properties are crucial for understanding complex medical scenarios, supporting early diagnosis and treatment decisions.

One of our feature plans is to validate the proposed method on larger and more heterogeneous real-world datasets, including multi-center medical records, imaging data, and time-series physiological signals. This will enable us to assess the method's robustness and generalizability across various clinical settings and data types. Additionally, we aim to tackle the critical challenges discussed in this paper, ensuring that the framework remains adaptable and efficient across diverse and complex environments.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine, 161*, 1-13. [CrossRef]

[2] Sun, B., & Wei, H. L. (2022, September). Machine learning for medical and healthcare data analysis and modelling: Case studies and performance comparisons of different methods. In *2022 27th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE. [CrossRef]

[3] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A, 379*(2194), 20200209. [CrossRef]

[4] Khojaste-Sarakhsi, M., Haghighi, S. S., Ghomi, S. F., & Marchiori, E. (2022). Deep learning for Alzheimer's disease diagnosis: A survey. *Artificial intelligence in medicine, 130*, 102332. [CrossRef]

[5] Chen, S. (1988). Representation of non-linear systems: the NARMAX model. *International Journal of Control, 49*(3), 303-344. [CrossRef]

[6] Billings, S. A. (2013). Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains. John Wiley & Sons.

[7] Nayyar, A., Gadhavi, L., & Zaman, N. (2021). Machine learning in healthcare: review, opportunities and challenges. *Machine learning and the internet of medical things in healthcare*, 23-45. [CrossRef]

[8] Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care, 48*(6), S106-S113. [CrossRef]

[9] Hosseini, M. P., Hosseini, A., & Ahi, K. (2020). A review on machine learning for EEG signal processing in bioengineering. *IEEE reviews in biomedical engineering, 14*, 204-218. [CrossRef]

[10] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley*

Interdisciplinary Reviews: Data Mining and Knowledge Discovery, *10*(5), e1379. [CrossRef]

[11] Jiang, K., Fujii, F., & Shiinoki, T. (2019). Prediction of lung tumor motion using nonlinear autoregressive model with exogenous input. *Physics in Medicine & Biology*, *64*(21), 21NT02. [CrossRef]

[12] Sum, J. P. F., Kan, W. K., & Young, G. H. (1999). A note on the equivalence of NARX and RNN. *Neural computing & applications*, *8*(1), 33-39. [CrossRef]

[13] Wei, H. L., & Billings, S. A. (2006). Feature subset selection and ranking for data dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, *29*(1), 162-166. [CrossRef]

[14] Pang, S., Meng, F., Wang, X., Wang, J., Song, T., Wang, X., & Cheng, X. (2020). VGG16-T: a novel deep convolutional neural network with boosting to identify pathological type of lung cancer in early stage by CT images. *International Journal of Computational Intelligence Systems*, *13*(1), 771-780. [CrossRef]

[15] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46. [CrossRef]

[16] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*, 1063-1095. [CrossRef]

[17] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91). [CrossRef]

[18] Palechor, F. M., & De la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief*, *25*, 104344. [CrossRef]

[19] UCI machine learning repository. (n.d.). UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/264/eeg+eye+state

[20] Vicchietti, M. L., Ramos, F. M., Betting, L. E., & Campanharo, A. S. (2023). Computational methods of EEG signals analysis for Alzheimer's disease classification. *Scientific Reports*, *13*(1), 8184. [CrossRef]

Bo Sun is a Ph.D. candidate in Data Science at the University of Sheffield. His research interests include interpretable machine learning, time-series modeling, and feature selection. He has conducted research on lightweight, transparent models in various real-world domains. His work enhances both the performance and interpretability of data-driven systems involving sequential and static data. (Email: bsun14@sheffield.ac.uk)

Hua-Liang Wei received the P.h.D. degree from the University of Sheffield in complex systems modelling, machine learning and information processing, Sheffield, Uk, in 2004. He is a Senior Lecturer (Associate Professor) at the University of Sheffield, where he leads research in signal processing and data-driven modeling of nonlinear complex systems, artificial intelligence, interpretable machine learning, and intelligent diagnosis, with applications in many multi-disciplinary domains. His research has been funded by EPSRC, NERC, STFC, the Royal Society, Horizon 2020 and other research councils. He has published more than 160 peer-reviewed papers. (Email: w.hualiang@sheffield.ac.uk)