ISSN: 2581-1274

A Benchmark of Activation Functions in Extreme Learning Machine for High-Dimensional Low-Sample-Size Microarray Classification

Inggih Permana [a,1], Shir Li Wang [a,2,*], Febi Nur Salisah [b,3], Sanusi [c,4], Febi Yanto [b,5]

[a] *Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Tanjong Malim, 35900, Malaysia*
[b] *Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia*
[c] *Teknologi Informasi, Fakultas Teknik, Universitas Teuku Umar*
[1]p20231001211@siswa.upsi.edu.my; [2]shirli_wang@meta.upsi.edu.my; [3]febinursalisah@uin-suska.ac.id; [4]sanusi@utu.ac.id; [5]febiyanto@uin-suska.ac.id*
**Corresponding author*

ARTICLE INFO

Article history:
Published
December 25, 2025

Keywords:
Activation function
Extreme learning machine
High-dimensional
Low-sample-size
Microarray

ABSTRACT

Microarray data classification presents significant challenges in machine learning due to its high-dimensionality, low-sample-size (HDLSS) characteristics, and class imbalance. These conditions often lead to overfitting and result in models with low performance stability. Extreme Learning Machine Learning (ELM) is an efficient method, but its performance is highly dependent on the choice of activation function that determines the model's non-linearity. Therefore, this study aims to conduct a comparative benchmarking of eight activation functions (sigmoid, tanh, relu, hard limit, identity, swish, cosine, and soft sign) within the ELM framework on 11 cancer-related microarray datasets. Testing was conducted through 30 independent runs to ensure statistical robustness, and performance was evaluated using accuracy, F1-score, precision, and recall metrics. Experimental results show that the choice of activation function has a significant impact on microarray data classification performance. The sigmoid function consistently provides superior and most stable results across various datasets, achieving a global average accuracy above 70% for all evaluation metrics. The advantage of the sigmoid lies not only in its high average performance but also in its stability, as evidenced by its low standard deviation. These findings provide strong practical guidance, recommending the use of the sigmoid activation function for robust ELM implementations in microarray data classification.

I. Introduction

Microarray technology serves as a crucial tool in genomics and bioinformatics, enabling the simultaneous measurement of expression levels of thousands of genes in biological samples, particularly for the diagnosis and classification of diseases such as cancer [1]. However, the resulting data is characterized by two major challenges: very high feature dimensionality and limited sample size. These characteristics of high-dimensionality, low-sample-size (HDLSS) inherently increase the risk of overfitting in machine learning models [2]. Furthermore, the class imbalance common in microarray data further complicates the classification challenge. This combination of challenges hinders the ability of conventional algorithms to identify relevant and generalizable biological patterns. Therefore, the development and application of efficient and robust classification methods, often through optimization of neural network architectures, are necessary to address the inherent complexity of such data.

Extreme Learning Machine Learning (ELM) is a learning algorithm developed for Single Hidden Layer Feedforward Networks (SLFN), with the primary goal of improving training efficiency and model generalization capabilities. In the ELM architecture, the connection weights between the input and hidden layers are randomly initialized, while the output weights are calculated analytically

through a least-squares solution [3]. The fundamental advantage of ELM lies in its exceptionally fast training speed, which significantly outperforms iteration-based methods such as backpropagation [4]. This effectiveness and efficiency have been proven in various applications, including gene expression classification in microarray data [5]. Due to its computational efficiency and generalization capabilities, ELM is a relevant choice for research in the field of bioinformatics and an ideal framework for comparative studies on the effects of activation functions.

Activation functions are essential components of the ELM architecture, as they play a critical role in nonlinearly transforming data from the input space to the hidden feature space. The choice of activation function directly affects the network's ability to represent and learn complex patterns in the data [6]. For example, functions such as tanh, rectified linear unit (relu), and sigmoid offer different nonlinearity capacities, which influence model performance based on the specific characteristics of the data being processed. Given the importance of this aspect, several studies, such as that by [6], have proposed the development of tunable activation functions to enhance the flexibility of ELM representations. This variation and optimization of activation functions underscores the fact that appropriate selection is crucial in maximizing the classification accuracy of microarray data. Therefore, comparative studies evaluating the impact of different activation functions are essential for understanding and optimizing the performance of ELM models in bioinformatics.

Although ELM has been widely applied in microarray data classification, a significant gap in the literature remains regarding the in-depth understanding of the influence of activation functions on model performance. ELM research on microarray data has traditionally focused on feature optimization and regularization parameters, as documented by [7]. However, a comprehensive comparative exploration of the contribution of various activation functions to model performance is still limited. The choice of activation function has critical implications for several model aspects, including convergence speed, generalization ability, and numerical stability of ELM. Previous studies, such as those by [8], have shown that modifying activation functions has the potential to improve feature representation and classification accuracy. However, a clear consensus has not yet been reached on the most optimal activation function for handling microarray data characterized by HDLSS and imbalanced properties. Therefore, this study aims to address this knowledge gap by systematically comparing various activation functions within the ELM framework. This effort is expected to provide new empirical insights into the causal relationship between activation function characteristics and biological classification effectiveness.

A study comparing activation functions in ELM has been conducted by [9]. However, the characteristics of the data used in this study show substantial differences compared to the study by [9], which classified active compounds based on the SMILES representation from PubChem. In that article, each chemical structure was extracted into 29 simple structural features, such as atom types, bonds, charges, and aromatic characteristics, resulting in a low-dimensional and relatively stable dataset for computational processing in ELM. In contrast, this study focused on microarray data that inherently has a very high number of features with limited samples, thus creating a HDLSS condition that demands a more comprehensive preprocessing and model optimization approach, including the selection of appropriate activation functions. Therefore, the level of data complexity in this study is much higher compared to the SMILES dataset in [9].

This study comprehensively explores eight activation functions: sigmoid, hyperbolic tangent (tanh), hardlimit, identity, swish, relu, cosine, and softsign. These functions were selected based on their relevance, empirical significance, and comparative role in the ELM literature. The sigmoid function was chosen due to its prevalence and proven superior performance in various previous ELM studies [6][9][10-13]. Similarly, the tanh function and its variants were considered due to their established use in comparative performance studies [9][13]. Among the more modern and popular functions, relu, frequently used in conventional deep learning architectures, was included to test its effectiveness within the ELM framework. The swish activation function was also used because it has been used as a comparison function in previous studies, such as [9]. Furthermore, softsign was included due to empirical evidence showing consistent and superior generalization performance

compared to other activation functions [14]. The cosine function was tested due to its role in several studies using ELM, such as [15]. Meanwhile, the Hardlimit function was retained as a binary comparative activation function; this function was also used by [10] and [9]. The identity function was used as a mandatory baseline. The inclusion of identity is crucial for a comprehensive evaluation, as it allows this study to explicitly measure and isolate the extent to which the improvement in feature mapping ability and classification accuracy of ELM is due to the non-linear transformation provided by the other activation functions.

The main objective of this study is to analyze and compare the performance of various activation functions in the ELM algorithm, specifically for microarray data classification tasks. Performance evaluation focuses on a series of performance measurement metrics including accuracy, precision, recall, and F1-score, as well as the stability of the results, which is very important in the context of HDLSS categorized data. Specifically, this study seeks to identify activation functions that are able to provide an optimal balance between feature representation complexity and robust prediction accuracy. It is hoped that the resulting findings can provide an empirical basis for the development of ELM models that are more adaptive and optimal to the unique characteristics of biological data, such as microarray gene expression data.

The contribution of this work is multi-fold. Different from previous comparative studies, this research systematically investigates activation functions within the ELM framework on HDLSS microarray data. This domain presents unique statistical and computational challenges compared to traditional low-dimensional datasets. Specifically, this study introduces a comprehensive empirical framework to evaluate eight different activation functions, ranging from classical to modern, and from linear to nonlinear, examining their varying effects on generalization performance, numerical stability, and classification robustness in the context of biological data. Unlike earlier research that has primarily focused on parameter optimization or feature selection, this study delves deeper into how activation functions influence the internal representational capacity of ELM under the constraints of HDLSS data. Furthermore, by incorporating a more diverse set of activation paradigms such as cosine and softsign, which have been largely overlooked in prior ELM research, this work provides new empirical evidence of their potential for modeling the complex nonlinear structures commonly observed in gene expression data. Thus, the core novelty of this work lies in bridging theoretical understanding with empirical analysis of activation dynamics in ELM for microarray data, laying a critical foundation for future development of adaptive and biologically informed learning architectures.

This article is structured into four main sections to ensure a logical and coherent flow of discussion. Section 2 details the applied methodology, including a description of the ELM algorithm, the activation functions explored, the characteristics of the dataset used, and the experimental setup procedure. Next, Section 3 presents the empirical findings from the experiments, followed by an in-depth analysis and discussion of the resulting performance. Finally, Section 4 summarizes the study's main contributions and provides a comprehensive conclusion. This structure is designed to facilitate the reader's understanding, from the conceptual foundations to the research's key achievements and implications.

II. Methodology

The ELM training procedure, based on the formulation proposed by Huang [16][17], can be outlined as follows. The training phase begins with defining the network architecture, which includes determining the number of neurons in the input, hidden, and output layers, as well as selecting an appropriate activation function. Subsequently, the input weight vectors (w) and biases are randomly initialized; these parameters remain static and are not updated during the learning process. The next step involves computing matrix H (the hidden layer output matrix), which represents the non-linear transformation of the data into a new feature space. This matrix H has dimensions where the rows correspond to the number of training samples and the columns correspond to the number of hidden neurons. The process concludes by calculating the output weights (β) using Equation 1.

$$\beta = H^{\dagger}T \tag{1}$$

In Equation 1, $H^{\dagger}$ denotes the Moore-Penrose pseudoinverse of matrix H, and T represents the target matrix containing the actual labels of the training data. Equation 2 is utilized to predict a new sample.

$$f(x) = h(x)\beta \tag{2}$$

In Equation 2, h(x) is the feature mapping vector comprising the set of output values from all nodes in the hidden layer for a given input sample x following transformation by the activation function.

Activation Functions

This subsection will present a detailed description of the eight activation functions explored in this study: sigmoid, tanh, hardlimit, identity, swish, relu, cosine, and softsign. The explanations will outline the mathematical characteristics and nonlinearity properties of each function, which are relevant in the context of implementing the ELM algorithm.

Sigmoid

The sigmoid activation function is one of the classical nonlinear functions most frequently used in ELM architectures. Its main characteristics are its smooth curve shape and its ability to map inputs into a limited range (0,1), making it suitable for generating controlled nonlinear mappings. In the studies of [9] and [10], sigmoid was used as one of the activation functions in their studies. Although sigmoid is known in the neural network literature to have a tendency to saturate at extreme values that can trigger vanishing gradients in backpropagation-based algorithms, this issue does not impact ELM because ELM does not rely on backpropagation and the output weights are calculated analytically. Furthermore, the sigmoid's simple mathematical form contributes to numerical stability in the formation of hidden layer matrices. With these characteristics, sigmoid remains a relevant basic activation function for use as a comparative benchmark in various ELM experiments. The sigmoid activation function formula is presented in Equation 3.

$$\text{sigmoid}(x) = 1 / (1 + e^{-x}) \tag{3}$$

Tanh

The tanh activation function is a nonlinear function with a similar curve shape to the sigmoid, but maps input values to the range (-1, 1). This property produces zero-centered outputs, which theoretically can improve training stability in artificial neural networks due to the balanced distribution of activations around zero. In the ELM framework, tanh is used in various studies as an alternative activation function to facilitate smooth nonlinear mappings [6][9][13]. Although tanh, similar to the sigmoid, is susceptible to saturation phenomena at extreme input values, it remains relevant for ELM experiments. This is due to its smooth, continuous, and bounded nature, which effectively supports the formation of robust nonlinear representations in the hidden layers. The tanh activation function formula is presented in Equation 4.

$$\tanh(x) = (e - e^{-x} / (e + e^{-x}) \tag{4}$$

Hardlimit

A hardlimit activation function is a piecewise binary function that produces discrete output values, typically 0 or 1. In the context of ELM research, it is often included as a comparative activation (baseline). For example, a study by [10] evaluated its performance alongside other activation functions in a proposed ELM variant. Similarly, [9] included a mathematically equivalent binary step function to the hardlimit as part of an extensive classification performance analysis. Despite its simplicity, the discrete nature of the hardlimit limits its ability to map and represent complex nonlinear patterns smoothly. Therefore, it generally serves as a comparative baseline to assess the contribution of nonlinearity, while smoother nonlinear activation functions tend to yield superior

performance on datasets with high complexity. The hardlimit activation function formula is presented in Equation 5.

$$\text{hardlimit}(x) = \{1, \text{if } x \geq 0;\ 0, \text{if } x < 0) \quad (5)$$

Identity

The identity activation function, or linear function, is the most basic activation function that maps input values directly to outputs without any nonlinear transformations. In neural network architectures, this function is commonly applied to the output layer for regression tasks, where the goal is to maintain a linear relationship between the input and the target variable [18]. However, in the hidden layers of ELM, the identity function does not meet the nonlinearity requirement essential for achieving universal approximation [17], and is therefore rarely used as the primary activation function. This function exhibits limitations in capturing complex data relationships because it inherently does not provide nonlinear transformations. Therefore, the identity function is used primarily as a comparative baseline. The goal is to explicitly quantify the extent to which the ELM performance improvement produced by nonlinear activation functions is significant compared to a purely linear model. The identity activation function formula is presented in Equation 6.

$$\text{identity}(x) = x \quad (6)$$

Swish

The Swish activation function, introduced by Google Brain [19], offers an adaptive blend of linear and nonlinear characteristics. It is known in modern neural network literature as a non-monotonic activation function with smooth gradients and is less susceptible to saturation than functions such as the sigmoid. Although the study by [9] does not explicitly discuss the gradient properties of Swish, they include Swish as one of the comparative activation functions within the ELM variant framework. In the context of ELM, the non-monotonic nature of Swish allows small negative values to still contribute to the non-linear representation, potentially increasing the flexibility of feature mapping. With these characteristics, Swish is a relevant and important modern alternative to evaluate in efforts to develop and improve ELM performance. The swish activation function formula is presented in Equation 7.

$$\text{swish}(x) = x.\text{sigmoid}(x) \quad (7)$$

Relu

The relu activation function is one of the most dominant activation functions in modern neural network architectures due to its simplicity and efficient computational advantages. In the context of ELM, relu is often included as a comparative activation function to evaluate the impact of different activation forms on model performance, as demonstrated by [9] in their extensive experiments. While relu in backpropagation-based networks is susceptible to the dying neuron problem (where neurons stop learning), this issue is not significant in ELM because the hidden layer weights are randomly initialized and not updated iteratively. Due to its lightweight computational characteristics and its ability to linearly map positive values, relu remains a relevant and important modern activation function to be used as a benchmark in ELM performance improvement studies. The relu activation function formula is presented in Equation 8.

$$\text{relu}(x) = \max(0, x) \quad (8)$$

Cosine

The cosine activation function is a nonlinear, periodic function used in the context of ELM to construct trigonometric wave-based representations. In a study by [15], cosine was evaluated as one of four activation functions, along with sigmoid, sine, and tanh, within their ELM variant framework. The authors included this function primarily as a comparison to the sine function, although cosine is not as popular as other activation functions in general practice. The periodic nature of cosine gives the model the potential to capture cyclic patterns in the data, although its effectiveness may be limited

if the data feature relationships do not exhibit a clear cyclic structure. Thus, cosine serves as a useful alternative activation function to evaluate the contribution of trigonometric patterns to the performance of ELM feature mapping. The cosine activation function formula is presented in Equation 9.

$$\text{cosine}(x) = \cos(x) \tag{9}$$

Softsign

The softsign activation function is a nonlinear alternative that offers smoother mapping compared to tanh, producing outputs in the range (−1, 1). In a study by [14], softsign was evaluated alongside various other activation functions within the ELM framework, and the results indicated that softsign consistently provides better generalization performance than many other activation functions. The smooth and less prone to saturation nature of softsign makes it effective in controlling the activation scale, which in turn can reduce the model's sensitivity to noise and improve the stability of the learning process. With these characteristics, softsign plays a relevant role as an activation function to strengthen the nonlinear mapping capabilities of ELM, especially on complex datasets that require stable representations that are resistant to activation saturation. The softsign activation function formula is presented in Equation 10.

$$\text{softsign}(x) = x / (1 + |x|) \tag{10}$$

Datasets

This study leverages a compilation of 11 publicly accessible microarray datasets, initially curated and utilized by [20]. These datasets are hosted and available for reproducibility at https://csse.szu.edu.cn/staff/zhuzx/datasets.html. A comprehensive summary characterizing the utilized data materials is provided in Table 1. Despite being established nearly two decades ago, the dataset collection compiled by [20] maintains significant relevance and wide adoption as a benchmark in contemporary research. This is evidenced by its frequent reappearance in recent literature; for instance, [21] employed a subset of four microarray datasets from this collection, while [22] similarly incorporated three of the original datasets for their experimental validation.

Table 1. Description of the microarray datasets used.

No.	Dataset	Total Genes	Instances	Classes
1	Leukemia - 2 Class	7129	72	2
2	Leukemia - 3 Class (A)	7129	72	3
3	Leukemia - 4 Class	7129	72	4
4	SRBCT	2308	83	4
5	Lung Cancer	12600	203	5
6	Ovarian Cancer	15154	253	2
7	Lymphoma	4026	66	3
8	Leukemia - 3 Class (B)	12582	72	3
9	Colon Tumor	2000	62	2
10	CNS	7129	60	2
11	Breast Cancer	24481	97	2

Experimental Setup

The experimental procedure was systematically designed for each combination of microarray dataset and activation function, as illustrated in Figure 1. A hold-out validation approach was used, where each dataset was divided into a training data (90%) and a testing data (10%).

The training data undergoes standardized pre-processing steps:

1. Missing value imputation: Handling missing values is done by filling in the missing values using the average (mean) value of the related features.
2. Feature selection: Feature selection is applied using the Information Gain metric to reduce dimensionality and increase feature relevance.
3. Data balancing: Synthetic Minority Over-sampling Technique (SMOTE) technique is used to address the class imbalance problem.
4. Data normalization: Data is normalized using Vector Normalization to equalize the feature scale.

After preprocessing the training data, modeling was performed using the ELM algorithm (with 64 hidden nodes) and eight activation functions (sigmoid, tanh, hardlimit, identity, swish, relu, cosine, and softsign). The performance of the eight models was then evaluated using the test data. It is important to note that the test data also underwent feature selection, data balancing, and normalization steps using identical transformation parameters derived from the training subset.

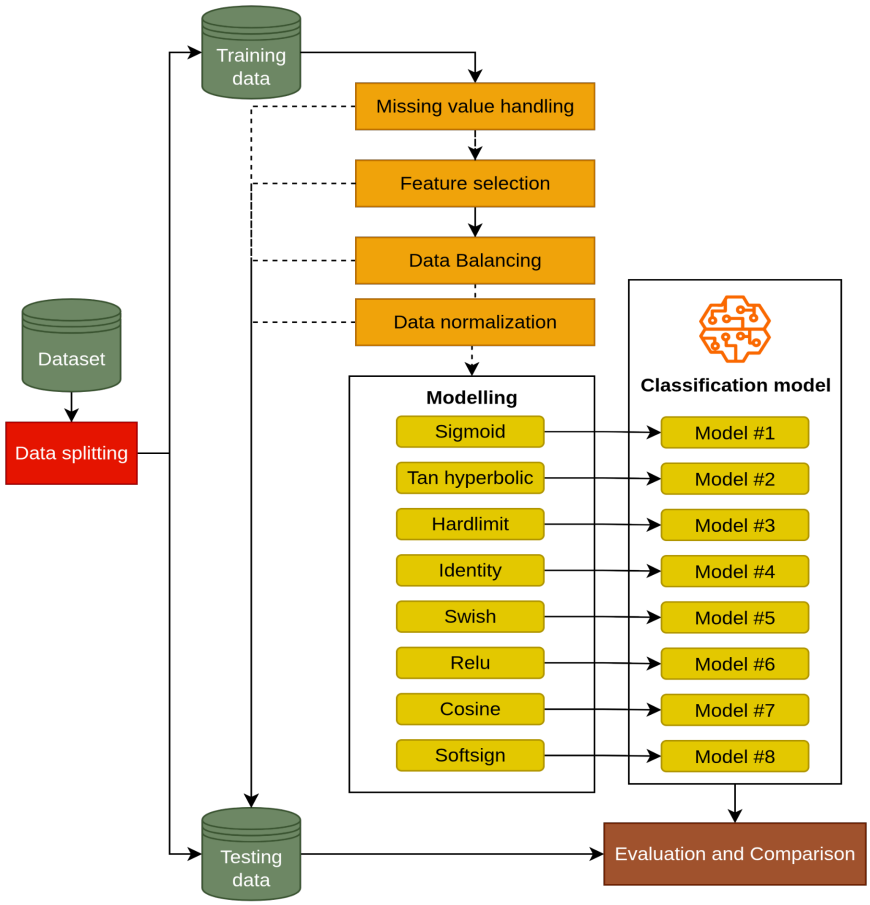

Fig. 1. Experimental procedure for each combination of a microarray dataset and activation function method.

Model performance was assessed using four main evaluation metrics: accuracy, precision, recall, and F1-score. To ensure statistical reliability and stability, each combination of dataset and activation function was run 30 times with different random seeds (1, 2, ..., 30). The final performance is reported as the average of the metrics used. Additionally, the standard deviation of each metric was calculated to provide a quantitative indicator of model performance stability. All experiments were implemented using the Python programming language on the Google Colab platform.

III. Results and Discussion

Table 1 presents the performance evaluation and ranking of specific activation functions per dataset in the context of ELM implementation. Meanwhile, Table 2 summarizes the average activation

function rankings calculated across all datasets used. The results of non-parametric statistical analysis using the Friedman test on these average rankings indicate statistically significant performance variations among the explored activation functions. Based on the comprehensive average rankings presented in Table 2, the sigmoid activation function was identified as having superior performance, followed in order by softsign, swish, relu, tan hyperbolic, identity, cosine, and finally hardlimit.

Table 1. Performance and ranking of activation functions per dataset.

Dataset	Activation Function	Accuracy		Precision		Recall		F1-Score	
		Score (%)	Rank	Score (%)	Rank	Score (%)	Rank	Score (%)	Rank
Breast Cancer	cosine	54.67	8	56.59	6	54.67	8	52.96	8
	hardLimit	58.67	3	59.03	3	58.67	3	56.5	4
	identity	57.33	5	58.15	4	57.33	5	56.33	5
	resume	**60**	**1**	60.28	2	**60**	**1**	**59.23**	**1**
	sigmoid	56.33	6	56.05	7	56.33	6	55.08	6
	softsign	57.67	4	58	5	57.67	4	56.64	3
	swish	59.67	2	**61.09**	**1**	59.67	2	58.7	2
	fishy	55.33	7	55.61	8	55.33	7	54.34	7
CNS	cosine	49.44	5	45.11	4	45.83	6	42.99	5
	hardLimit	43.89	8	42.08	8	42.5	8	39.4	8
	identity	51.67	4	43.72	6	46.25	4.5	43.74	4
	resume	48.33	6	43.47	7	44.58	7	41.82	7
	sigmoid	52.22	2.5	49.08	2	49.58	3	48.4	2
	softsign	52.22	2.5	**50.94**	**1**	50.83	2	**48.53**	**1**
	swish	**55**	**1**	48.5	3	**51.25**	**1**	48.29	3
	fishy	46.67	7	44.42	5	46.25	4.5	42.02	6
Colon Tumor	cosine	70	3	67.61	2	**70**	**1.5**	66.49	3
	hardLimit	61.9	8	63.22	7	64.33	7	58.99	7
	identity	63.81	7	59.22	8	59.67	8	58.78	8
	resume	68.57	4	66.08	4	68.5	3	65.21	4
	sigmoid	**72.86**	**1.5**	67.08	3	68	4	67.2	2
	softsign	68.1	5	64.5	5	65.67	6	63.57	5
	swish	**72.86**	**1.5**	**68.25**	**1**	**70**	**1.5**	**68.11**	**1**
	fishy	67.62	6	64.25	6	65.83	5	63.31	6
Leukemia – 2 Class	cosine	81.67	7	81.12	7	81.33	7	80.09	7
	hardLimit	70.42	8	72.16	8	71	8	68.85	8
	identity	86.67	5	87.17	5	87.56	5	86.03	5
	resume	87.5	4	87.94	4	88.44	4	87.08	4
	sigmoid	**94.17**	**1**	**94.31**	**1**	**95.11**	**1**	**94.01**	**1**
	softsign	90.83	2	91.36	2	91.11	2	90.23	2
	swish	85.83	6	85.61	6	86.67	6	85.33	6
Leukemia – 3 Class (A)	fishy	88.33	3	88.92	3	89.56	3	87.97	3
	cosine	52.5	8	46.23	8	48.89	8	45.57	8
	hardLimit	57.92	6	54.39	2	53.43	5	50.98	6
	identity	57.92	6	**55.35**	**1**	**59.07**	**1**	**55.15**	**1**
	resume	61.67	3	53.7	4	55.37	2	52.18	3
	sigmoid	62.5	2	52.83	6	53.89	4	51.99	4
	softsign	57.92	6	50.95	7	52.59	7	50.15	7
	swish	**62.92**	**1**	53.34	5	55.28	3	52.24	2
	fishy	58.75	4	54.32	3	53.33	6	51.45	5
Leukemia – 4 Class	cosine	62.08	7	50.86	6	55.21	7	50.16	6
	hardLimit	54.17	8	50.35	7	59.38	6	49.86	7
	identity	62.92	6	48.53	8	53.96	8	49.12	8
	resume	72.92	3	67.44	3	70.42	2	66.17	3
	sigmoid	**80.83**	**1**	**70.04**	**1**	**74.79**	**1**	**70.52**	**1**
	softsign	76.67	2	67.47	2	70	3	67.36	2
	swish	71.25	4	62.65	4	66.04	4	61.72	4
	fishy	70.83	5	59.24	5	65	5	59.43	5

Dataset	Activation Function	Accuracy		Precision		Recall		F1-Score	
		Score (%)	Rank	Score (%)	Rank	Score (%)	Rank	Score (%)	Rank
Lung Cancer	cosine	84.13	7	74.65	7	70.9	7	71.01	7
	hardLimit	61.11	8	56.71	8	69.19	8	56.87	8
	identity	**91.43**	**1**	79.4	4	76.57	4	77.04	4
	resume	87.3	6	79.56	3	**80.38**	**1**	78.71	2
	sigmoid	91.27	2	80.87	2	77.81	3	78.31	3
	softsign	90.63	3	**81.17**	**1**	78.9	2	**78.96**	**1**
	swish	87.94	4	77.24	5	74.9	5	74.66	5
	fishy	87.78	5	76.22	6	73.81	6	73.42	6
Lymphoma	cosine	75.24	4	71.69	2	83.11	2	71.92	2
	hardLimit	**89.05**	**1**	**84.85**	**1**	**90.44**	**1**	**85.42**	**1**
	identity	61.43	8	51.65	8	59.78	8	48.72	8
	resume	77.14	2	70.46	3	80.44	3	70.76	3
	sigmoid	74.76	5	69.37	5	77.56	6	69.1	5
	softsign	75.71	3	70.19	4	78	5	69.49	4
	swish	70	7	62.83	7	74.44	7	62.87	7
	fishy	72.86	6	68.7	6	79.33	4	68.17	6
Leukemia – 3 Class (B)	cosine	63.33	7	63.93	6	61.48	7	59.61	7
	hardLimit	59.58	8	56.09	8	56.3	8	54.38	8
	identity	64.58	5.5	63.11	7	62.22	5	59.88	6
	resume	64.58	5.5	64.22	5	62.04	6	61.04	5
	sigmoid	**75.83**	**1**	**76.33**	**1**	**73.7**	**1**	**73.09**	**1**
	softsign	70.83	2	73.65	2	68.89	2	68.2	2
	swish	65.83	4	66.91	4	63.89	4	62.39	4
	fishy	67.08	3	71.67	3	65.74	3	65.29	3
Ovarian Cancer	cosine	97.18	6	96.42	6	97.84	6	96.98	6
	hardLimit	79.74	8	80.53	8	82.94	8	79.33	8
	identity	97.69	2	97.08	2	98.15	2	97.51	2
	resume	96.15	7	95.38	7	97.06	7	95.92	7
	sigmoid	**97.82**	**1**	**97.41**	**1**	**98.25**	**1**	**97.67**	**1**
	softsign	97.31	5	96.64	5	97.94	5	97.12	5
	swish	97.56	3.5	97.03	4	98.05	4	97.39	4
	fishy	97.56	3.5	97.07	3	98.14	3	97.42	3
SRBCT	cosine	49.26	6	47.16	7	50.14	7	45.79	7
	hardLimit	51.85	4	52.97	4	54.44	4	49.55	4
	identity	55.56	2	54.17	3	58.33	3	55.95	2
	resume	46.67	8	48.56	5	48.61	8	45.19	8
	sigmoid	**63.33**	**1**	**67.99**	**1**	**69.03**	**1**	**64.71**	**1**
	softsign	54.44	3	57.83	2	60.83	2	55.77	3
	swish	51.11	5	46.15	8	51.67	6	45.89	6
	fishy	48.89	7	47.46	6	52.78	5	47.41	5

Table 2. Aggregated rank of activation functions across datasets.

Activation Function	Rank				
	Accuracy	Precision	Recall	F1-Score	Average
sigmoid	**2.18**	**2.73**	**2.82**	**2.46**	**2.55**
softsign	3.41	3.27	3.64	3.18	3.37
swish	3.55	4.36	3.96	4.00	3.90
resume	4.50	4.27	4.00	4.27	4.34
fishy	5.14	4.91	4.68	5.00	4.87
identity	4.68	5.09	4.86	4.82	4.91
cosine	6.18	5.55	6.05	6.00	5.96
hardLimit	6.36	5.82	6.00	6.27	6.11

Based on the aggregation of rankings across datasets (summarized in Table 2), the sigmoid activation function consistently demonstrated superior performance, achieving the best average ranking across all key evaluation metrics, namely accuracy, precision, recall, and F1-score. This superiority was validated by the global mean performance analysis (Figure 2), where the sigmoid function dominated with the highest global mean values across all tested metrics. Quantitatively, the sigmoid function achieved a global average of 74.72% for accuracy, 71.03% for precision, 72.19% for recall, and 70.01% for F1-score. This stable and superior performance across various metrics and datasets confirms the substantial generalization capability of the sigmoid function on microarray classification tasks using the ELM architecture.

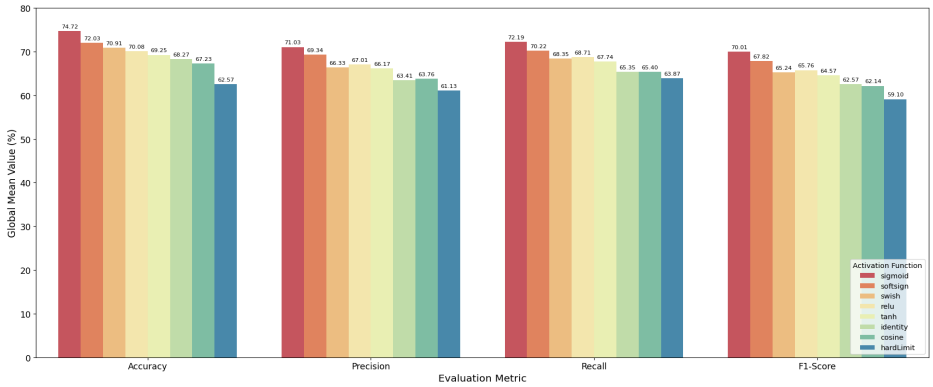

Fig. 2. Global mean performance of ELM activation functions across all datasets.

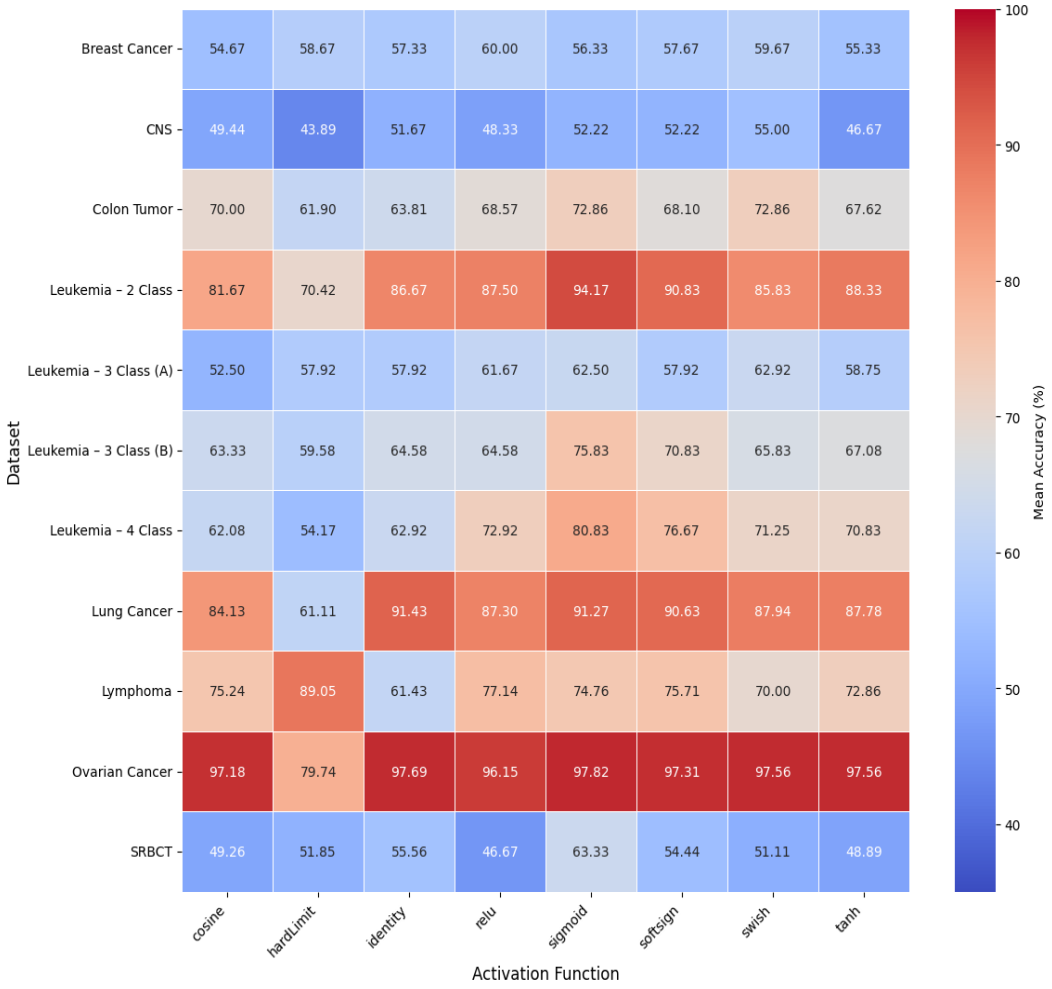

Fig. 3. Heatmap of mean accuracy for ELM across datasets and activation functions.

The softsign, swish, and relu activation functions ranked second, third, and fourth, respectively, in this test. These three functions showed generally superior average performance compared to the tanh, identity, cosine, and hardlimit functions, although their performance remained below the highest performance achieved by sigmoid. This observation is supported by the global average performance data, where softsign, swish, and relu recorded global mean accuracy above 70.00%. Specifically, only softsign was able to achieve a global mean recall value above 70.00% among the three functions, while none reached the 70.00% threshold for precision and F1-Score.

Although swish ranked third, a detailed analysis of each dataset (as visualized in Figure 3, Heatmap) revealed significant potential as an alternative. Specifically, swish achieved the best accuracy among all activation functions on the Leukemia - 3 Class (A), Colon Tumor, and CNS datasets. Another interesting fact is that relu, despite being a very popular and dominant activation function in conventional Deep Neural Network architectures, failed to show any notable superiority in the context of microarray data classification using this ELM Algorithm.

In contrast, the tanh, identity, cosine, and hardlimit functions produced average accuracies below the overall average of the experiments, a trend clearly visible in the scatterplot of the data (Figure 3). The hardlimit function consistently recorded the worst global average performance across all metrics used. While there was an anomaly where hardlimit managed to outperform all metrics on the Leukemia - 3 Class (B) dataset, this was limited to only one dataset and does not reflect strong overall generalization capabilities.

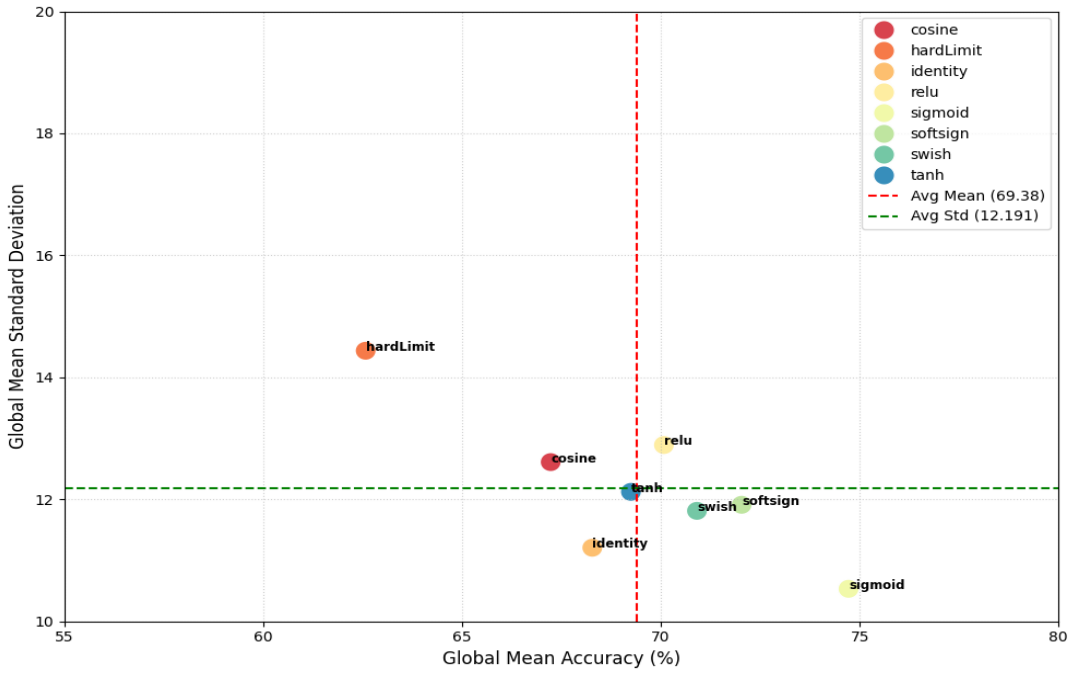

Fig. 4. Comparison between standard deviation and global mean accuracy of activation functions.

Figure 4 presents a scatter plot comparing the average global accuracy with the average global standard deviation of the activation functions, which serves as a visual indicator of performance stability. In this analysis, the ideal activation function lies in the high accuracy and low standard deviation quadrant. The sigmoid activation function reaches the optimal position, characterized by the highest average global accuracy and the lowest average global standard deviation. These results confirm that the ELM model using sigmoid not only produces superior performance but is also highly stable and robust to random initialization variations. Furthermore, the swish and softsign functions show a promising combination, as both functions successfully record average global accuracy above the overall mean while maintaining an average global standard deviation below the overall mean. This phenomenon indicates that both functions provide relatively consistent performance improvements.

Conversely, the relu function exhibits a global mean accuracy above the overall mean but also has a global average standard deviation that is also above the overall mean. This indicates that while relu

has the potential to produce good solutions, its performance tends to be unstable and sensitive to initial initialization. Meanwhile, identity and tanh show good stability (standard deviations below the overall mean), but are limited by global mean accuracy that is below the overall mean. The worst performance is shown by cosine and hardlimit, both of which produce below-average accuracy and have high standard deviations. Specifically, hardlimit is at the worst position, confirming its low and highly unstable performance.

In summary, these results have important practical implications for microarray classification and the broader field of bioinformatics. The consistent superiority of the sigmoid activation function across various evaluation metrics highlights its ability to produce stable and generalizable feature mappings from high-dimensional microarray data. Specifically, this suggests that ELM models utilizing sigmoid activation are highly suitable for diagnostic and predictive tasks where data dimensionality and sample sparsity are critical concerns, such as cancer subtype classification, biomarker discovery, and patient stratification. Additionally, the strong performance of softsign and swish supports their potential use in adaptive or hybrid ELM frameworks within bioinformatics pipelines that require both high accuracy and robustness to noisy gene expression profiles. From a computational standpoint, the high stability of sigmoid and softsign enables more efficient parameter tuning and improved model reproducibility, which are essential in clinical bioinformatics workflows. Overall, these findings offer practical guidance for selecting and optimizing activation functions in ELM-based bioinformatics applications, supporting the development of faster, more interpretable, and biologically meaningful classification systems.

IV. Conclusion

This study concludes that the selection of activation function plays a crucial role in determining the performance of the ELM algorithm when applied to microarray classification tasks. Through comprehensive evaluation on 11 HDLSS microarray datasets, the sigmoid activation function consistently outperformed others across multiple performance metrics, including accuracy, precision, recall, and F1-score. Its superiority is not only reflected in average performance but also in its numerical stability, as indicated by a low standard deviation across datasets. These findings suggest that sigmoid-based ELM architectures are highly suitable for bioinformatics tasks such as cancer subtype classification, biomarker discovery, and patient stratification, where data sparsity and non-linearity are major challenges.

In contrast, the ReLU activation function, despite its widespread success in deep learning, failed to deliver competitive results in the ELM framework for microarray data. This indicates that activation functions commonly used in deep neural networks may not directly transfer to ELM architectures in the HDLSS context. Similarly, the hardlimit function demonstrated the poorest performance, characterized by both low accuracy and high variability, and is therefore not recommended for similar tasks.

Despite the robust insights offered by this study, several limitations must be acknowledged. First, the scope of this work is limited to only 11 microarray datasets, which, although diverse, may not fully represent the variability found across different types of genomic data or other omics domains. Second, the evaluation focused solely on single-layer ELM architectures, without exploring deeper or ensemble-based ELM variants, which may behave differently under various activation functions. Third, the study did not consider the impact of preprocessing techniques, such as feature selection or normalization strategies, which could interact with activation function behavior.

Future research can address these limitations by expanding the dataset pool to include a wider range of biological data types, including RNA-Seq or epigenomic profiles. Investigating the interplay between activation functions and advanced ELM variants, such as deep ELM or kernel-based ELM, would also provide a more comprehensive understanding. Moreover, integrating adaptive or hybrid activation strategies, possibly leveraging data-driven selection mechanisms, could further improve classification performance and model interpretability in complex bioinformatics pipelines.

References

[1] BEGUM, Shemim, et al. Application of active learning in DNA microarray data for cancerous gene identification. Expert Systems with Applications, 2021, 177: 114914. https://doi.org/10.1016/j.eswa.2021.114914.

[2] VOGIATZIS, Dimitrios; TSAPATSOULIS, Nicolas. Active learning for microarray data. International Journal of Approximate Reasoning, 2008, 47.1: 85-96. https://doi.org/10.1016/j.ijar.2007.03.009.

[3] ZHANG, Runxuan, et al. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. IEEE/ACM transactions on Computational Biology and Bioinformatics, 2007, 4.3: 485-495. https://doi.org/10.1109/tcbb.2007.1012.

[4] KUMAR, C. Arun; RAMAKRISHNAN, S. Binary classification of cancer microarray gene expression data using extreme learning machines. In: 2014 IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2014. p. 1-4. https://doi.org/10.1109/ICCIC.2014.7238297.

[5] SUCHARITA, Swati, et al. Classification of cancer microarray data using a two-step feature selection framework with moth-flame optimization and extreme learning machine. Multimedia Tools and Applications, 2024, 83.7: 21319-21346. https://doi.org/10.1007/s11042-023-16353-2.

[6] LI, Bin; LI, Yibin; RONG, Xuewen. The extreme learning machine learning algorithm with tunable activation function. Neural Computing and Applications, 2013, 22.3: 531-539. https://doi.org/10.1007/s00521-012-0858-9.

[7] HIRA, Swati; BAI, Anita. A novel map reduced based parallel feature selection and extreme learning for micro array cancer data classification. Wireless Personal Communications, 2022, 123.2: 1483-1505. https://doi.org/10.1007/s11277-021-09196-3.

[8] CAO, Jiuwen, et al. Extreme learning machine with affine transformation inputs in an activation function. IEEE transactions on neural networks and learning systems, 2018, 30.7: 2093-2107. https://doi.org/10.1109/TNNLS.2018.2877468.

[9] RATNAWATI, Dian Eka, et al. Comparison of activation function on extreme learning machine (ELM) performance for classifying the active compound. In: AIP Conference Proceedings. AIP Publishing LLC, 2020. p. 140001. https://doi.org/10.1063/5.0023872.

[10] ISMAIL, Noraini; OTHMAN, Zulaiha Ali; SAMSUDIN, Noor Azah. Regularization activation function for extreme learning machine. International Journal of Advanced Computer Science and Applications, 2019, 10.3: 240-247. https://dx.doi.org/10.14569/IJACSA.2019.0100331.

[11] LIU, Shenglan, et al. Robust activation function and its application: Semi-supervised kernel extreme learning method. Neurocomputing, 2014, 144: 318-328. https://doi.org/10.1016/j.neucom.2014.04.041.

[12] TRIPATHI, Diwakar, et al. Evolutionary extreme learning machine with novel activation function for credit scoring. Engineering Applications of Artificial Intelligence, 2020, 96: 103980. https://doi.org/10.1016/j.engappai.2020.103980.

[13] ALIMI, Fathi, et al. M-estimation activation functions for high-performance extreme learning machine ensemble classification. Scientific Reports, 2025, 15.1: 32154. https://doi.org/10.1038/s41598-025-16798-5.

[14] DAS, Dibyasundar, et al. An empirical evaluation of extreme learning machine: application to handwritten character recognition. Multimedia Tools and Applications, 2019, 78.14: 19495-19523. https://doi.org/10.1007/s11042-019-7330-0.

[15] DOKEROGLU, Tansel; SEVINC, Ender. Evolutionary parallel extreme learning machines for the data classification problem. Computers & Industrial Engineering, 2019, 130: 237-249. https://doi.org/10.1016/j.cie.2019.02.024.

[16] HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541). Ieee, 2004. p. 985-990. https://doi.org/10.1109/IJCNN.2004.1380068.

[17] HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. Extreme learning machine: theory and applications. Neurocomputing, 2006, 70.1-3: 489-501. https://doi.org/10.1016/j.neucom.2005.12.126.

[18] HASTIE, Trevor, et al. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 2005, 27.2: 83-85. https://doi.org/10.1007/978-0-387-84858-7.
[19] RAMACHANDRAN, Prajit; ZOPH, Barret; LE, Quoc V. Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941, 2017, 7.1: 5.
[20] ZHU, Zexuan; ONG, Yew-Soon; DASH, Manoranjan. Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition, 2007, 40.11: 3236-3248. https://doi.org/10.1016/j.patcog.2007.02.007.
[21] DENG, Xiongshi, et al. Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. Medical & Biological Engineering & Computing, 2022, 60.3: 663-681. https://doi.org/10.1007/s11517-021-02476-x.
[22] SALISAH, Febi Nur, et al. Evaluating the Impact of Data Balancing Techniques on the k-Nearest Neighbors Algorithm for Microarray Data Classification. Jurnal Inotera, 2025, 10.2: 261-271. https://doi.org/10.31572/inotera.Vol10.Iss2.2025.ID497.