



Multi-Kernel Fusion for RBF Neural Networks

Syed Muhammad Atif¹ · Shujaat Khan² · Imran Naseem^{3,4} · Roberto Togneri⁴ · Mohammed Bennamoun⁵

Accepted: 8 June 2022 / Published online: 24 June 2022
© The Author(s) 2022

Abstract

A simple yet effective architectural design of radial basis function neural networks (RBFNN) makes them amongst the most popular conventional neural networks. The current generation of radial basis function neural network is equipped with multiple kernels which provide significant performance benefits compared to the previous generation using only a single kernel. In existing multi-kernel RBF algorithms, multi-kernel is formed by the convex combination of the base/primary kernels. In this paper, we propose a novel multi-kernel RBFNN in which every base kernel has its own (local) weight. This novel flexibility in the network provides better performance such as faster convergence rate, better local minima and resilience against sticking in poor local minima. These performance gains are achieved at a competitive computational complexity compared to the contemporary multi-kernel RBF algorithms. The proposed algorithm is thoroughly analysed for performance gain using mathematical and graphical illustrations and also evaluated on three different types of problems namely: (i) pattern classification, (ii) system identification and (iii) function approximation. Empirical results clearly show the superiority of the proposed algorithm compared to the existing state-of-the-art multi-kernel approaches.

Keywords Pattern classification · Function approximation · Non-linear system identification · Neural networks · Radial basis function · Gaussian kernel · Support vector machine · Euclidean distance · Cosine distance · Kernel fusion

1 Introduction

Machine learning (ML) is an established field with a wide range of applications including control engineering [5, 18, 24, 29], medical imaging [23, 35, 47], bioinformatics [26, 31, 41], and design of forecasting systems [11, 19, 36, 48], etc. It has been successfully used for other innovative applications as well such as in the design of cognitive communication systems [6, 34] and powerful generative models for number of multimedia application [13, 27]. In

Syed Muhammad Atif, Shujaat Khan: Both authors contributed equally.

✉ Imran Naseem
imran.naseem@uwa.edu.au

Extended author information available on the last page of the article

ML, neural networks are considered to be an important category of tools being frequently used. Therefore number of neural network architectures for example spiking neural neural network (SPNN), multiple layer perceptron (MLP), convolutional neural networks (CNN) and radial basis function neural network (RBFNN) has been proposed.

Due to its compact design and good noise tolerance RBFNN is extensively used in various applications where computational complexity, and data availability is a constrain [4]. Several advances have been proposed to improve its performance. For instance, to improve the parameter learning a variant of gradient decent has been proposed [24], instead of gradient descent algorithms some researchers have used meta-heuristic algorithms to update kernel weights and other network parameters [3, 4, 39, 46]. Aljarah *et al.* in [4], used bio-geography-based optimization algorithm (BBO) [39]. Alexandridis *et al.* studied the effectiveness of particle swarm algorithm (PSO) for updating weights of the RBFNN [3].

Recently researchers have successfully blended RBFNN with other established techniques as well. For example [28, 44, 45], Yang et al. in [45] proposed an efficient method for the selection of the centers using the conventional K-means clustering. However, unnecessary points around cluster centers were removed during global K-means clustering using population density method. This slight tweak in the selection procedure of the center, resulted in faster convergence and more robustness. In [44], Wena et al. used Takagi-Sugeno (TS) fuzzy model with the RBF neural network. The proposed designed is particularly useful in environments with data loss, data distortion or signal saturation. It uses K-means clustering for both selecting fuzzy rules and the centers of the RBFNN. Moreover, weighted activation degree (WAD) is used to determine the firing strength of fuzzy node. Liu et al. [28] proposed C-RBFNN (Cloud RBFNN) which uses the cloud theory in fuzzy mathematics to optimize the activation functions. This modification allows RBFNN to effectively express the fuzziness and randomness of the user data such as social media data.

Some hybrid training options have also been recently explored. For instance in [8], Yao and Kuo proposed to combine self-organizing map (SOM) based RBF with evolutionary algorithms such as partial swarm optimization (PSO) and genetic algorithm (GA). This hybrid approach for RBF outperformed conventional non-hybrid approaches. Another emerging variant of RBFNN called spatio-temporal RBFNN, uses the concept of time-space orthogonality to separately model the dynamics and nonlinear complexities [20, 36]. Additionally, an adaptive Nelder Mead Simplex [12], based training method that simultaneously updates weights and kernel width is proposed in [15].

1.1 Motivation and Contribution of this Research

RBFNN typically uses a single type of kernel lacking better generalization. This is because practical learning problems often involve multiple, heterogeneous data sources. Hence, the choice of kernel is heavily dependent on the problem at hand [1, 10]. For example, wavelet kernel, due to its excellent local properties both in time and frequency domains, performs better for some signal approximation and pattern classification problems, however due to lack of prior knowledge choosing the best kernel for the given learning problem is a challenging task. An alternative approach is to use multiple kernels to incorporate design flexibility and generalization [7, 10, 42]. This approach has been successfully employed with other kernel-based methods for instance in support vector machine (SVM) [40, 43]. The most widely used approach to combine multiple kernels of different characteristics is convex combination i.e. all participating kernels are combine linearly such that their coefficients are non-negative and sum to unity [30, 40, 43]. Recently, some researchers have made successful attempts to

combine multiple kernels in a nonlinear fashion e.g. Gu, Yanfeng, et al. in [14] showed the effectiveness of combining multiple kernels using Hadamard product.

In the context of RBFNN, multi kernel approach is still an under-explored research area. Fu et al. [10] were the first to introduce the multi kernel RBF-NN. They combined the Gaussian kernel and the wavelet kernel using convex combination and adaptively tuned the kernel coefficients using orthogonal least squares (OLS) algorithm. Later, Aftab *et al.* in [1] and Khan *et al.* in [25] explored the area of multi-kernel RBFNN and designed an adaptive multi-kernel RBFNN. Motivated from these works, we propose a novel multi-kernel RBFNN architecture as a Coordinating RBF Neural Network (Co-RBFNN).

Conventional multi-kernel RBF architectures, use the concept of linear combination of various primary kernels (Gaussian, cosine, wavelet etc) with either fixed or adaptive weights, incorporating single degree of freedom [1, 10, 25]. In particular, the conservative choice of the mixing parameters turns out to be the limitation of these conventional approaches. In contrast, the proposed kernel fusion method uses matrix-based mixing weights allowing each participating kernel to learn independently, thereby yielding better performance in most cases. This learning approach of independent mixing weights, make our method novel and unique compared to other contemporary approaches. The main contributions of our research are as follows:

1. A multi-kernel RBFNN architecture is proposed that combines each multi-kernel in the network with its own set of kernel parameters (local weights).
2. Graphical explanation of the algorithm is given to conceptually justify the origin of improved performance.
3. A comprehensive mathematical analysis is performed to identify the convergence bound.
4. The proposed architecture is evaluated for three problems of estimation namely non-linear system identification, pattern classification, and function approximation and extensive comparative analysis is performed with the contemporary approaches.

The organization of the paper is as follows. In Sect. 2, a brief overview of existing multi-kernel RBFNNs is proposed followed by the proposed Co-RBFNN in Sect. 3. Experimental evaluation and comparative results are discussed in Sect. 4. Finally, the paper is concluded in Sect. 5.

2 Multi-Kernel Radial Basis Function Neural Networks

2.1 Overview of the Architecture of the RBF Neural Network

RBFNN is a simple feed forward neural network that consists of only three layers i.e., an input layer, a nonlinear hidden layer and a linear output layer. Fig. 1 depicts the architecture of an RBFNN. Let $\mathbf{X} \in \mathbb{R}^{a \times S}$ representing an input dataset consist of S samples, and $\mathbf{x}_s \in \mathbb{R}^{a \times 1}$ be the input vector representing a sample by a number of attributes, then the overall mapping of the RBF network, $f: \mathbb{R}^{a \times 1} \rightarrow \mathbb{R}^{1 \times 1}$, is given as:

$$y_s = \sum_{k=1}^K w_k \phi_k(\mathbf{x}_s, \mathbf{m}_k) + b, \quad (1)$$

where for all k , $\mathbf{m}_k \in \mathbf{M} \in \mathbb{R}^{a \times K}$, K is the number of neurons in the hidden layer of the network, $\mathbf{M} \in \mathbb{R}^{a \times K}$ comprises of K number of $\mathbf{m}_k \in \mathbb{R}^{a \times 1}$ vectors, each representing a center point of the kernel of k th hidden neuron, w_k is the synaptic weight connecting the

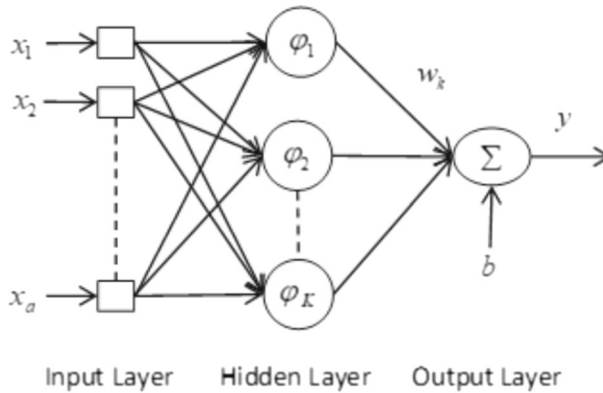


Fig. 1 Architecture of the RBF neural network

k th hidden neuron to the output neuron, b is the bias term of the output neuron and ϕ_k is the radial basis function of the k th hidden neuron. Without the loss of generality and for the sake of simplicity a single output neuron is considered. Conventional RBF networks employ a number of kernels such as multiquadrics, inverse multiquadrics and Gaussian [16].

2.2 Overview of the Contemporary Multi-Kernel Approaches

Gaussian kernel is considered to be the most commonly used kernel:

$$\phi_g(\mathbf{x}, \mathbf{m}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{m}\|^2}{\sigma^2}\right), \quad (2)$$

where σ is the kernel-width of the Gaussian kernel.

Recently, it has been argued that the cosine kernel offers complimentary information compared to the Gaussian kernel [1]. It is defined as:

$$\phi_c(\mathbf{x}, \mathbf{m}) = \frac{\mathbf{x} \cdot \mathbf{m}}{\|\mathbf{x}\| \|\mathbf{m}\| + \epsilon}, \quad (3)$$

where, $\|\cdot\|$ is the L2 norm or Euclidean distance and $\epsilon > 0$ is a small constant added to avoid the indeterminant form of Eq (3).

In recent studies [7, 14, 40, 42], it is suggested that combining multiple kernels is more efficient than using the kernels individually. Accordingly, a novel multi-kernel has been proposed combining cosine and Gaussian kernels [1]:

$$\phi_k(\mathbf{x}, \mathbf{m}_k) = \alpha_g \phi_g(\mathbf{x}, \mathbf{m}_k) + \alpha_m \phi_c(\mathbf{x}, \mathbf{m}_k), \quad (4)$$

where $\phi_g(\mathbf{x}, \mathbf{m}_k)$ and $\phi_c(\mathbf{x}, \mathbf{m}_k)$ are output of Gaussian and cosine kernels for k th hidden neuron respectively and, α_g and α_c are their corresponding kernel weights. Further, there are two constraints on α_g and α_c , i.e., $0 \leq \alpha_g, \alpha_c \leq 1$ and $\alpha_g + \alpha_c = 1$. The common set of kernel weights i.e., $\{\alpha_g, \alpha_c\}$ for all multi-kernels and the above two constraints ensures that the participating kernels will form a convex combination.

The new multi-kernel in (4) has shown some good results compared to the conventional Gaussian kernel [1]. In this method, the fusion of the two kernels is manual and the their weights α_g and α_c are adjusted in a hit-and-trial manner. Without any prior information, a

common practice is to assign equal weights to the two kernels i.e. $\alpha_g = \alpha_c = 0.5$. To resolve this issue, in [25], an adaptive framework is proposed for automatic fusion of kernels. This approach tunes the kernel weights at every iteration n to minimize error [25]:

$$\phi_k(\mathbf{x}, \mathbf{m}_k) = \alpha_g(n)\phi_g(\mathbf{x}, \mathbf{m}_k) + \alpha_c(n)\phi_c(\mathbf{x}, \mathbf{m}_k). \quad (5)$$

In [25], both the synaptic weights of hidden neuron and kernel weights are updated using the conventional gradient descent algorithm. This method has shown improvement over the fixed multi-kernel methods [1].

3 The Proposed Coordinating RBFNN (Co-RBFNN)

Motivated by [25], we argue that this adaptive scheme can be further improved by introducing a separate set of kernel weights for each participating kernel. Therefore, the k th kernel of the given RBFNN that consists of two participating kernels will take the form:

$$\phi_k(\mathbf{x}, \mathbf{m}_k) = \alpha_{g_k}(n)\phi_g(\mathbf{x}, \mathbf{m}_k) + \alpha_{c_k}(n)\phi_c(\mathbf{x}, \mathbf{m}_k), \quad (6)$$

where $\phi_{g_k}(\mathbf{x}, \mathbf{m}_k)$ and $\phi_c(\mathbf{x}, \mathbf{m}_k)$ are the Gaussian and cosine contributors of the k th multi-kernel with the corresponding weights $\alpha_{g_k}(n)$ and $\alpha_{c_k}(n)$ respectively. Eq (6) can be rewritten as:

$$\phi_k(\mathbf{x}, \mathbf{m}_k) = \sum_l^L \alpha_{l_k}(n)\phi_{l_k}(\mathbf{x}, \mathbf{m}_k), \quad (7)$$

where, $l \in L$ and $L = \{g, c\}$ is the set of participating primary kernels in the k th multi-kernel. So, ϕ_{l_k} is the l th participating primary kernel of the k th kernel and α_{l_k} is its mixing weight.

Eq (7) can be easily extended for more than two kernels. However, we restrict ourselves to only two kernels for the sake of simplicity. The overall mapping at the n th iteration can be written as:

$$y(n) = \sum_{k=1}^K w_k(n) \left(\sum_{l \in \{g, c\}} \alpha_{l_k}(n)\phi_{l_k}(\mathbf{x}(n), \mathbf{m}_k) \right) + b(n), \quad (8)$$

where K is the number of centers (multi-kernel) of the network, $\mathbf{m}_k \in \mathbb{R}^{a \times 1}$ is the center of the k th multi-kernel, w_k is the synaptic weight connecting the k th hidden neuron to the output neuron, b is the bias term of the output neuron, ϕ_{l_k} is the l th participating kernel of k th multi-kernel and α_{l_k} is the corresponding kernel weight.

Eq. (8) can be written as:

$$\begin{aligned} y(n) &= \sum_{k,l} \left(w_k(n)\alpha_{l_k}(n) \right) \phi_{l_k}(\mathbf{x}(n), \mathbf{m}_k) + b(n) \\ &= \sum_{k,l} w_{k,l}(n)\phi_{l_k}(\mathbf{x}(n), \mathbf{m}_k) + b(n), \end{aligned} \quad (9)$$

where, $k = 1, 2, \dots, K$, $l \in \{g, c\}$ and $w_{k,l}(n) = w_k(n)\alpha_{l_k}(n)$ is the substitute form of the weight of l th participating kernel in the k th multi-kernel. $\mathbf{x}(n)$ is a sample obtained from \mathbf{X} at n th iteration.

It is evident from Eq (9) that there is no explicit need to maintain kernel weight of each participating kernel of a given multi-kernel. Instead, each participating kernel ϕ_{l_k} has

its own corresponding weight $w_{k,l}(n)$. In other words, our proposed multi-kernel RBFNN architecture, consisting of K hidden neurons and L participating kernels (in our case $L = 2$), may be unfolded into a simple RBFNN architecture consisting of $K \times L$ centers (hidden neurons), such that there are L sets of K hidden neurons and each of that set employs one of the L different kernels.

In matrix form, Eq (9) can be written as:

$$y(n) = \boldsymbol{\phi}^T(n) \mathbf{w}(n), \quad (10)$$

where, $\mathbf{w}(n) = [b, w_{g_1}(n), w_{g_2}(n), \dots, w_{g_K}(n), w_{c_1}(n), w_{c_2}(n), \dots, w_{c_K}(n)]^T$ and $\boldsymbol{\phi}(n) = [1, \phi_{g_1}(\mathbf{x}(n), \mathbf{m}_k), \dots, \phi_{g_K}(\mathbf{x}(n), \mathbf{m}_k), \phi_{c_1}(\mathbf{x}(n), \mathbf{m}_k), \dots, \phi_{c_K}(\mathbf{x}(n), \mathbf{m}_k)]^T$ are weights and kernel vectors respectively and $[\cdot]^T$ is the vector transpose operation.

3.1 Weight and Bias Update Rules

The update rule of the synaptic weight $w_{k,l}(n)$ at $(n + 1)$ th iteration can be given as:

$$w_{k,l}(n + 1) = w_{k,l}(n) + \Delta w_{k,l}(n), \quad (11)$$

$$\Delta w_{k,l}(n) = -\eta \frac{\partial \ell}{\partial w_{k,l}(n)}, \quad (12)$$

where, η is the learning rate, and ℓ is the mean-square-error (L2) loss function defined as:

$$\ell(\mathbf{w}, b) = \frac{1}{N} \sum_{n=1}^N (d(n) - y(n))^2. \quad (13)$$

The above loss function can be minimized by solving for the instantaneous error, considering instantaneous error function $\mathcal{E}(n)$ i.e.,:

$$\mathcal{E}(n) = \mathcal{E}(\mathbf{w}(n), b(n)) = \frac{1}{2} (d(n) - y(n))^2, \quad (14)$$

where $d(n)$ is the desired output, $y(n)$ is the actual output at the n th iteration and $e(n)$ the instantaneous error.

Using the chain rule of differentiation for the cost function in Eq (14) yields:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{k,l}(n)} = \frac{\partial \mathcal{E}(n)}{\partial e(n)} \frac{\partial e(n)}{\partial y(n)} \frac{\partial y(n)}{\partial w_{k,l}(n)}, \quad (15)$$

which upon simplification of the partial derivatives in Eq (15) results in:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{k,l}(n)} = -e(n) \phi_{l_k}(\mathbf{x}(n), \mathbf{m}_k). \quad (16)$$

Using Eq (12) and Eq (16), the update rule in Eq (11) will becomes:

$$w_{k,l}(n + 1) = w_{k,l}(n) + \eta e(n) \phi_{l_k}(\mathbf{x}(n), \mathbf{m}_k), \quad (17)$$

similarly, the update rule for bias $b(n)$ can be shown to have the form:

$$b(n + 1) = b(n) + \eta e(n). \quad (18)$$

3.2 Training Algorithm

For the training of the proposed network, the steps of the algorithm outlined in Table 1 are followed. Define the inputs, $X \in \mathbb{R}^{a \times S}$, $M \in \mathbb{R}^{a \times K}$ (where the columns are the centers of the K multi-kernels) the initial weight matrix $W_{init} \in \mathbb{R}^{K \times L}$, initial value of bias b , the learning rate $\eta > 0$ and T number of epochs for training. The algorithm yields a weight matrix $W \in \mathbb{R}^{K \times L}$ as output. Conventional stochastic gradient descent is used to update the weight matrix $W \in \mathbb{R}^{K \times L}$ independently using each of the S training samples in each of the T epochs.

Table 1 Algorithmic depiction of the proposed Co-RBFNN.

Require: An a -by- S training data matrix X consist of S samples of a dimension (features/attributes), an 1-by- S training desired response matrix d of corresponding S samples, K -by- L kernel functions Φ , a -by- K matrix M for K multi-kernel centers (means), an K by L initial weight matrix W_{init} , initial bias b_{init} , η the learning rate for the weights and bias, and T number of training epochs.

Ensure: an K by L final weight matrix W

Initialize: $W = W^{(prev)} = W_{init}$; $b = b^{(prev)} = b_{init}$; $t = 1$;

repeat

$s = 1$;

repeat

$k = 1$; $y_s = b$;

repeat

$l = 1$;

repeat

$y_s = y_s + \mathbf{w}_{k,l} \phi_{l_k}(\mathbf{x}_s, \mathbf{m}_k)$;

$l = l + 1$;

until $l \leq L$;

$k = k + 1$;

until $k \leq K$;

$e_s = d_s - y_s$;

$W^{(prev)} = W$; $b^{(prev)} = b$;

$k = 1$;

repeat

$l = 1$;

repeat

$\mathbf{w}_{k,l} = \mathbf{w}_{k,l}^{(prev)} + \eta e_s \phi_{l_k}(\mathbf{x}_s, \mathbf{m}_k)$;

$l = l + 1$;

until $l \leq L$;

$k = k + 1$;

until $k \leq K$;

$b = b^{(prev)} + \eta e_s$;

$s = s + 1$;

until $s \leq S$

$t = t + 1$;

until $t \leq T$

3.3 Illustrative Explanation of the Proposed Method

In this subsection, we consider an illustrative example depicted in Fig. 2. The task is to classify a test point. It is illustratively proved that a primary kernel (which is a Gaussian or a cosine kernel in this example) fails to effectively discriminate the given test point. In contrast, our proposed solution effectively maps the given test point to its true class. This illustration therefore serve to demonstrate the superiority of the proposed method. For the purpose of this illustrative case-study, no assumptions were made except the choice of a highly challenging test point to prove the efficacy of the proposed algorithm for difficult cases.

As depicted in Fig. 2, we consider a challenging binary classification problem, in which the only tunable parameters are the kernel mixing weights. We have four center points obtained using a clustering method such as K-mean clustering (or any other method) representing two classes namely *ClassA* and *ClassB*. As shown in Fig. 2, *Center1_A* and *Center2_A* are the representative points of *ClassA* and *Center1_B* and *Center2_B* are the representative points of *ClassB* respectively. Let's consider a test sample *TestPoint_A* such that *dc1_A*, *dc2_A* are Euclidean distances from *TestPoint_A* to centers *Center1_A* and *Center2_A* respectively whereas *dc1_B*, *dc2_B* are Euclidean distances of test sample *TestPoint_A* from centers *Center1_B* and *Center2_B* respectively. Similarly, *ac1_A*, *ac2_A* are angles of test sample *TestPoint_A* with centers *Center1_A* and *Center2_A* respectively whereas *ac1_B*, *ac2_B* are angles of test sample *TestPoint_A* with centers *Center1_B* and *Center2_B* respectively.

Without loss of generality, weights of the model are set to unity. Now, the following relationships hold on model at the time of presentation of test sample *TestPoint_A*.

$$dc1_A = dc2_B, \quad (19)$$

$$dc2_A = dc1_B, \quad (20)$$

$$ac1_A > ac1_B > ac2_B > ac2_A, \quad (21)$$

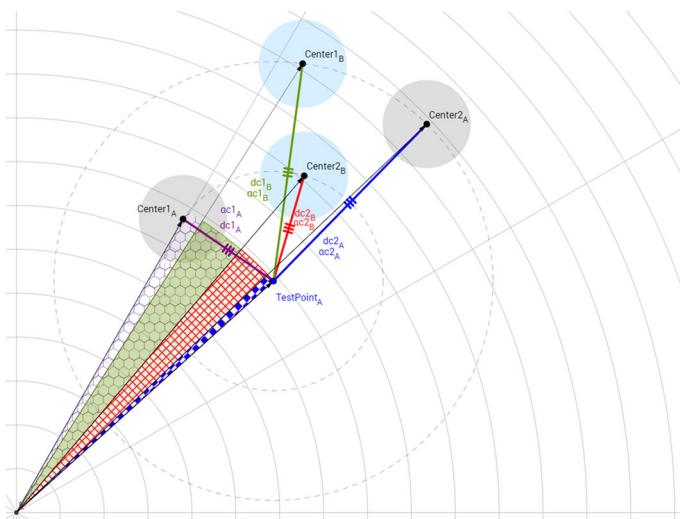


Fig. 2 Illustrative explanation of the proposed RBF algorithm

$$\begin{aligned} & \phi_c(\text{Test Point}_A, \text{Center1}_A) + \phi_c(\text{Test Point}_A, \text{Center2}_A) \\ &= \phi_c(\text{Test Point}_A, \text{Center1}_B) + \phi_c(\text{Test Point}_A, \text{Center2}_B). \end{aligned} \quad (22)$$

Let Ψ is the discriminative power of a classifier. For Gaussian and cosine kernel classifier, their discriminative powers are respectively equivalent to:

$$\begin{aligned} \Psi_g &= \phi_g(\text{Test Point}_A, \text{Center1}_A) + \phi_g(\text{Test Point}_A, \text{Center2}_A) \\ &\quad - (\phi_g(\text{Test Point}_A, \text{Center1}_B) + \phi_g(\text{Test Point}_A, \text{Center2}_B)), \end{aligned} \quad (23)$$

and

$$\begin{aligned} \Psi_c &= \phi_c(\text{Test Point}_A, \text{Center1}_A) + \phi_c(\text{Test Point}_A, \text{Center2}_A) \\ &\quad - (\phi_c(\text{Test Point}_A, \text{Center1}_B) + \phi_c(\text{Test Point}_A, \text{Center2}_B)). \end{aligned} \quad (24)$$

Using (19) and (20), we get:

$$\Psi_g = 0, \quad (25)$$

similarly, using (21) and (22), we get:

$$\Psi_c = 0. \quad (26)$$

Since, both Ψ_g and Ψ_c are zero the probability that Test Point_A belong to Class A is equal to that of Class B i.e. equiprobable using either Gaussian or cosine classifier. The classification of Test Point_A is therefore solely dependent on the value of the bias.

This lacking of correctly classifying a challenging cases such as Test Point_A persists even in RBF networks equipped with adaptive kernel fusion (Khan et al.) having global kernel weights as its discriminating power Ψ_a for (Khan et al.) is defined as:

$$\Psi_a = \alpha_g \Psi_g + \alpha_c \Psi_c, \quad (27)$$

where $\alpha_g \in \mathbb{R}$ and $\alpha_c \in \mathbb{R}$ are (global) kernel coefficients of Gaussian and cosine kernels respectively.

Again for difficult cases such as Test Point_A , it is verifiable that $\Psi_a = 0$

In contrast, the proposed method is not susceptible to such problems due to the novel concept of local weights (kernel coefficient) of each kernel. The discriminative power Ψ_r of Co-RBFNN can be written as:

$$\begin{aligned} \Psi_r &= \alpha_{\text{Center1}_A, g} \phi_g(\text{Test Point}_A, \text{Center1}_A) \\ &\quad + \alpha_{\text{Center2}_A, g} \phi_g(\text{Test Point}_A, \text{Center2}_A) \\ &\quad + \alpha_{\text{Center1}_A, c} \phi_c(\text{Test Point}_A, \text{Center1}_A) \\ &\quad + \alpha_{\text{Center2}_A, c} \phi_c(\text{Test Point}_A, \text{Center2}_A) \\ &\quad - \{ \alpha_{\text{Center1}_B, g} \phi_g(\text{Test Point}_A, \text{Center1}_B) \\ &\quad + \alpha_{\text{Center2}_B, g} \phi_g(\text{Test Point}_A, \text{Center2}_B) \\ &\quad + \alpha_{\text{Center1}_B, c} \phi_c(\text{Test Point}_A, \text{Center1}_B) \\ &\quad + \alpha_{\text{Center2}_B, c} \phi_c(\text{Test Point}_A, \text{Center2}_B) \}, \end{aligned} \quad (28)$$

where $\alpha_{c,x} \in \mathbb{R}$ is the kernel coefficient for kernel of type x and center c such that $x \in g, c$ and $c \in \text{Center1}_A, \text{Center2}_A, \text{Center1}_B, \text{Center2}_B$

It is evident that $\Psi_r \neq 0$ as $\alpha_{\text{Center1}_A, g} \neq \alpha_{\text{Center2}_A, g}$, $\alpha_{\text{Center1}_A, c} \neq \alpha_{\text{Center2}_A, c}$, $\alpha_{\text{Center1}_B, g} \neq \alpha_{\text{Center2}_B, g}$ and $\alpha_{\text{Center1}_B, c} \neq \alpha_{\text{Center2}_B, c}$ in general.

3.4 Mean Convergence Analysis of Our Proposed Model

In this subsection, we mathematically prove that our proposed algorithm will effectively converge provided that we strategically set the learning rate η less than λ_{max} , the maximum eigenvalue of the auto-correlation matrix R . We assume that, for the Wiener filter, the signal and (additive) noise are stationary linear stochastic processes with known spectral characteristics or known auto-correlation and cross-correlation [17].

The weight update rules of our proposed model i.e. (17) and (18) in the matrix form can be collectively rewritten as:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta \boldsymbol{\phi}(n)e(n), \quad (29)$$

where η is the learning rate, $\mathbf{w}(n)$ is the weight vector of n th iteration and e is the error between the desired and actual output signals i.e.

$$e(n) = d(n) - y(n). \quad (30)$$

Let's define the vector $\boldsymbol{\Delta}_{opt}$ as the difference of our proposed model estimated weight vector $\mathbf{w}(n)$ with the optimal weight vector \mathbf{w}_{opt} :

$$\boldsymbol{\Delta}_{opt}(n) = \mathbf{w}(n) - \mathbf{w}_{opt}, \quad (31)$$

where optimal weight vector \mathbf{w}_{opt} is that of Wiener filter obtained by solving the standard equation of Wiener filter i.e.

$$\mathbf{P} - \mathbf{R}\mathbf{w}_{opt} = 0, \quad (32)$$

where \mathbf{P} is the cross-correlation matrix between input signal to m hidden neurons (i.e. $\boldsymbol{\phi}$) and desired output d , and \mathbf{R} is the auto-correlation matrix of input signal to m hidden neurons i.e. $\boldsymbol{\phi}$. Mathematically,

$$\mathbf{R} = E(\boldsymbol{\phi}(n)\boldsymbol{\phi}^T(n)), \quad (33)$$

$$\mathbf{P} = E(\boldsymbol{\phi}(n)d). \quad (34)$$

Substituting the value of e from (30) and subtracting \mathbf{w}_{opt} from both sides of (29), we get:

$$\boldsymbol{\Delta}_{opt}(n+1) = \boldsymbol{\Delta}_{opt}(n) + \eta \boldsymbol{\phi}(n)(d - y(n)). \quad (35)$$

Substituting the value of y and $\mathbf{w}(n)$ from (10) and (31) respectively into (29), we get:

$$\boldsymbol{\Delta}_{opt}(n+1) = \boldsymbol{\Delta}_{opt}(n) + \eta \boldsymbol{\phi}(n)(d - \boldsymbol{\phi}^T(n)(\mathbf{w}_{opt} + \boldsymbol{\Delta}_{opt}(n))). \quad (36)$$

Taking expectation on both sides of (36) and rearranging few term, we obtain:

$$\begin{aligned} E(\boldsymbol{\Delta}_{opt}(n+1)) &= E(\boldsymbol{\Delta}_{opt}(n)) + \eta E(\boldsymbol{\phi}(n)d) \\ &\quad - \eta E(\boldsymbol{\phi}(n)\boldsymbol{\phi}^T(n)(\mathbf{w}_{opt} + \boldsymbol{\Delta}_{opt}(n))). \end{aligned} \quad (37)$$

Further simplifying the above equation using (32), (33) and (34), we get:

$$E(\boldsymbol{\Delta}_{opt}(n+1)) = E(\boldsymbol{\Delta}_{opt}(n)) - \eta E(\boldsymbol{\phi}(n)\boldsymbol{\phi}^T(n)\boldsymbol{\Delta}_{opt}(n)), \quad (38)$$

After applying usual assumptions of Wiener filter [17], we obtain:

$$E(\boldsymbol{\Delta}_{opt}(n+1)) = (I - \eta R)E(\boldsymbol{\Delta}_{opt}(n)). \quad (39)$$

Decomposing R using singular value decomposition (SVD) and further simplification leads us to:

$$0 < \eta < \frac{1}{\lambda_{\max}}, \quad (40)$$

where, λ_{\max} is the maximum eigenvalue of the autocorrelation matrix R .

3.5 Mathematical Analysis of the Proposed Model Co-RBFNN

In this subsection, we mathematically prove that our proposed solution is superior to the adaptive kernel fusion [25]. We prove that the mean square error of our proposed solution is always less than that of the adaptive kernel fusion [25]. During this mathematical analysis, we made a usual assumption that the errors induced by the two models (i.e. our proposed solution and adaptive kernel fusion [25]) are zero mean Gaussian noise.¹

Lemma 1 *Our proposed model has following relationship with adaptive kernel fusion (Khan et al.) model [25]*

$$y_d = y_a + e_x, \quad (41)$$

where, y_d and y_a are the estimated responses of our proposed model and adaptive kernel fusion [25] respectively and e_x is the noise. Mathematically, the estimated responses of the two models y_a and y_d respectively are defined as:

$$y_a = \alpha \mathbf{w}^T \boldsymbol{\phi}_g + (1 - \alpha) \mathbf{w}^T \boldsymbol{\phi}_c, \quad (42)$$

and,

$$y_d = \mathbf{w}_g^T \boldsymbol{\phi}_g(\mathbf{x}) + \mathbf{w}_c^T \boldsymbol{\phi}_c(\mathbf{x}), \quad (43)$$

where \mathbf{w}_g and \mathbf{w}_c are Gaussian and cosine weight vectors of our proposed model respectively and, \mathbf{w} and α are the weight vector and multi-kernel coefficient of adaptive kernel fusion [25] respectively.

Prove: Consider our proposed model that estimates the desired response by minimizing the least square error i.e.

$$d = y_d + e, \quad (44)$$

where, d is the desired response vector, y_d is the estimated response of our proposed model and $e \in \mathcal{N}(0, \sigma)$ is the Gaussian noise of the proposed model.

Further, the following relationships hold among weight vectors \mathbf{w} , \mathbf{w}_g and \mathbf{w}_c :

$$\mathbf{w}_g = \alpha \mathbf{w} + \mathbf{e}_g, \quad (45)$$

$$\mathbf{w}_c = (1 - \alpha) \mathbf{w} + \mathbf{e}_c, \quad (46)$$

where $\mathbf{e}_g \in \mathcal{N}(0, \sigma_g)$ and $\mathbf{e}_c \in \mathcal{N}(0, \sigma_c)$ are Gaussian noises and α is the kernel coefficient of multi-kernel as defined in adaptive kernel fusion [25].

By adding (45) and (46), we get another relation i.e.

$$\mathbf{w}_g + \mathbf{w}_c = \mathbf{w} + \mathbf{e}_g + \mathbf{e}_c. \quad (47)$$

Adding and subtracting the term $\mathbf{w}_g^T \boldsymbol{\phi}_c(\mathbf{x})$ on R.H.S of (41), substituting the value of y_d from (43) and simplifying, we get:

$$d = \mathbf{w}_g^T (\boldsymbol{\phi}_g(\mathbf{x}) - \boldsymbol{\phi}_c(\mathbf{x})) + (\mathbf{w}_g + \mathbf{w}_c)^T \boldsymbol{\phi}_c(\mathbf{x}) + e. \quad (48)$$

¹ Without loss of generality, the bias of the considered RBF models are assumed to be zero during the proofs of the following lemma and its two corollaries.

After substituting the value of \mathbf{w}_g from (45) and that of $(\mathbf{w}_g + \mathbf{w}_c)$ from (47) into (48) and simplifying, we obtain:

$$d = \alpha \mathbf{w}^T \boldsymbol{\phi}_g + (1 - \alpha) \mathbf{w}^T \boldsymbol{\phi}_c + \mathbf{e}_g^T \boldsymbol{\phi}_g(\mathbf{x}) + \mathbf{e}_c^T \boldsymbol{\phi}_c(\mathbf{x}) + e. \quad (49)$$

After substituting the value of $\alpha \mathbf{w}^T \boldsymbol{\phi}_g + (1 - \alpha) \mathbf{w}^T \boldsymbol{\phi}_c$ from (42), we obtain:

$$d = y_a + \mathbf{e}_g^T \boldsymbol{\phi}_g(\mathbf{x}) + \mathbf{e}_c^T \boldsymbol{\phi}_c(\mathbf{x}) + e. \quad (50)$$

Let the error term $\mathbf{e}_g^T \boldsymbol{\phi}_g(\mathbf{x}) + \mathbf{e}_c^T \boldsymbol{\phi}_c(\mathbf{x})$ be represented as e_x , (50) becomes:

$$d = y_a + e_x + e, \quad (51)$$

substituting the value of d from (44) into (51) and simplifying, we get:

$$y_d = y_a + e_x, \quad \text{Q.E.D} \quad (52)$$

Corollary 1 *The error term e_x is mean zero Gaussian noise i.e. $e_x \in \mathcal{N}(0, \sigma_x)$.*

Prove: Since adaptive kernel fusion [25] estimates the desired response d by minimizing the least square error. Therefore, it is mathematically definable as:

$$d = y_a + e_a, \quad (53)$$

where, y_a is the estimated response and $e_a \in \mathcal{N}(0, \sigma_a)$ is the Gaussian noise of the model respectively and d is the desired response vector.

Substituting the value of d from (51) into (53) and simplifying, we get:

$$e_x = e_a - e. \quad (54)$$

Since, e_x is the difference of two zero mean Gaussian noises i.e. e and e_a , e_x is also a zero mean Gaussian noise i.e. $e_x \in \mathcal{N}(0, \sigma_x)$, hence proved.

Corollary 2 *Mean squared error of adaptive kernel fusion (Khan et al.) model [25] $\|e_a\|_2^2$ is always greater than or equal to that of our proposed model $\|e_x\|_2^2$ i.e.*

$$\|e_a\|_2^2 \geq \|e\|_2^2. \quad (55)$$

Prove: Substituting the value of d from (51) into (53) and simplifying, we get:

$$e_a = e + e_x, \quad (56)$$

Since, $e_a \in \mathcal{N}(0, \sigma_a)$ is the sum of two mean zero Gaussian noises i.e. $e \in \mathcal{N}(0, \sigma)$ and $e_x \in \mathcal{N}(0, \sigma_x)$. Hence,

$$\sigma_a^2 = \sigma^2 + \sigma_x^2. \quad (57)$$

This lead us to:

$$\|e_a\|_2^2 = \|e\|_2^2 + \|e_x\|_2^2,$$

so,

$$\|e_a\|_2^2 \geq \|e\|_2^2,$$

hence, proved.

4 Experimental Results

In this section, we compare the performance of our proposed solution against two state-of-the-art multi-kernel radial basis function neural network algorithms namely manually fused multi-kernel proposed by Aftab et al. [1] and adaptively fused multi-kernel proposed by Khan et al. in [25]. All three algorithms are tested on pattern classification, system identification and function approximation problems for standard performance measures. All tests are preformed using Matlab R2017b on Intel CORE i5-2540M CPU @2.60GHz 4GB RAM. Results are averaged over 100 independent random runs.

4.1 Pattern Classification

Pattern classification has several applications in security, industry, medicine and defense. Examples include iris identification, speaker identification, fingerprint identification, statistical pattern recognition of seismic data, and automatic medical diagnosis.

A well known Iris flower dataset [9] is selected for pattern classification problem. The dataset consist of three classes (flower species). Each class has 50 samples and four attributes i.e. sepal length, sepal width, petal length, and petal width. Forty samples of each class are randomly selected for training where as remaining ten samples of each class are used for testing.

The three RBF networks are trained with the following specifications. 16 neurons are used with kernel centers selected using subtractive clustering [33] with influence factor 0.2. Gaussian kernel width is set to unity. Learning rate is 5×10^{-3} . The weights as well as bias are initialized randomly.

Fig. 3 shows MSE curves obtained during training. It is evident that our proposed architecture requires only 160 epochs to achieve mean squared error of -30.17 dB whereas the other two algorithms require at least 240 epochs to reach the same MSE. Moreover, the proposed architecture settles on an MSE of -35.39 dB after 2000 epoch whereas the other two algorithms achieve a worse error of -33.33 dB after same number of epochs. Hence, our proposed architecture outperforms other two state-of-the-art techniques both in term of rate of convergence and steady-state error.

Classification accuracy achieved by different RBF algorithms on the given dataset is shown in Table 1. During the training phase, the proposed architecture showed accuracy of 98.35% that is 0.64% higher than that manual kernel fusion [1] but 0.24% less compared to the adaptive kernel fusion [25] that attain the accuracy of 98.59%. However, our proposed approach attained the best testing accuracy of 99.13% comparing to 97.00% that of manual kernel fusion [1] and 98.50% that of adaptive kernel fusion [25]. It established that the proposed architecture is significantly tolerable to over-fitting. Moreover, our architecture is even not susceptible to the initial weights (and the bias) as it exhibited the lowest standard deviation of 0.12% (on the training data) and the second lowest standard deviation of 1.47% (on the test data). Fig. 4 and Fig. 5 show the training and testing accuracy curves of the three approaches respectively. Our proposed architecture exhibited better training accuracy from the start thus achieved the training accuracy of 95.67% at 100 epoch whereas the other two algorithm achieved 92.84% only at 100 epoch. On testing data, the manual kernel fusion [1] initially exhibited the best accuracy precisely 96.5% at 100. But, our proposed approach became the best at 600 epoch and marked the best steady-state accuracy of 99.27% at 2000 epoch comparing to that 98.27% by adaptive kernel fusion [25] and 97.23% by manual kernel fusion [1].

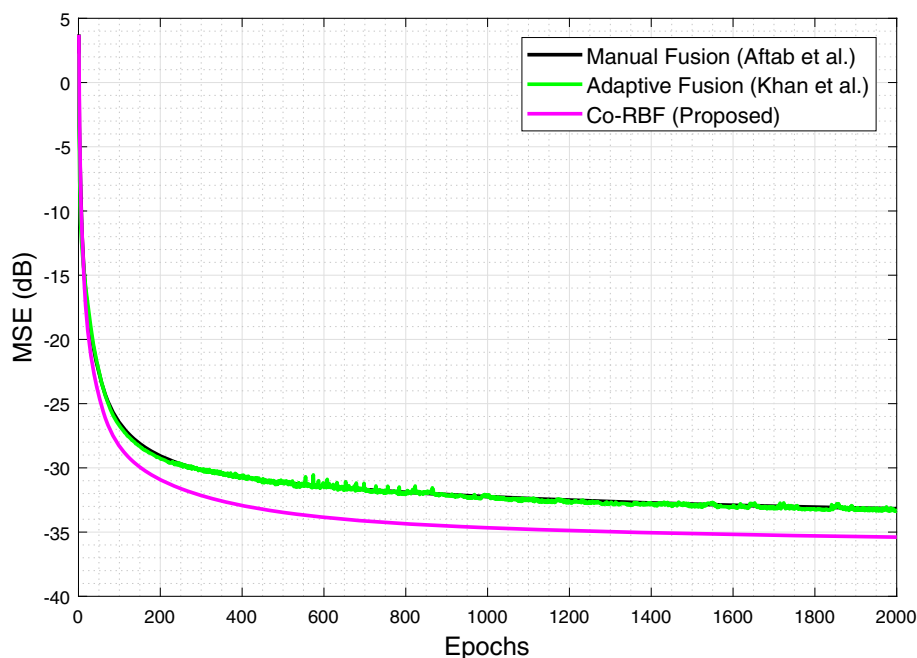


Fig. 3 MSE curves of different RBF algorithms on Iris Flowers dataset

Table 1 Classification accuracy (in %) of Iris Flowers dataset obtained by different RBF algorithms

| Method | Training mean \pm std | Testing mean \pm std |
|-------------------------------|----------------------------------|----------------------------------|
| Manual Fusion (Aftab et al.) | 97.71 \pm 0.61 | 97.00 \pm 1.01 |
| Adaptive Fusion (Khan et al.) | 98.59\pm1.12 | 98.50 \pm 4.68 |
| Co-RBF (Proposed) | 98.35 \pm 0.12 | 99.13\pm1.47 |

Sensitivity and specificity are also two important performance metric to analyze a classifier for its biasedness of a classifier. Sensitivity and specificity of different algorithms are tabulated in Table 2 and Table 3 respectively. Our proposed algorithm exhibits the best sensitivity of 97.50% and 100% on Versicolor and Setosa classes respectively during training and that of 100% and 100% on Virginica and Versicolor classes respectively in testing phases. Moreover, the sensitivity obtained by the proposed algorithm for all three classes are very close to each other in the range of 0% to 0.35% in testing phase showing unbiasedness of the proposed method.

During the training phase, our proposed algorithm shows the best specificity of 98.75% and 100% on Versicolor and Setosa classes respectively. Whereas, it achieved the average specificity of 98.75 on Versicolor class which is the second best specificity (i.e. 0.55% less than that of the best specificity of 99.33% reached by adaptive kernel fusion [25]) on that class. Specificity results of testing phase are also very similar. Our algorithm attained the specificity of 100% on both Versicolor and Setosa classes. However, it achieved the specificity of 98.70% on Versicolor class which is the second best specificity on that class, 0.35% less than the best (99.05%) attained by adaptive kernel fusion [25].

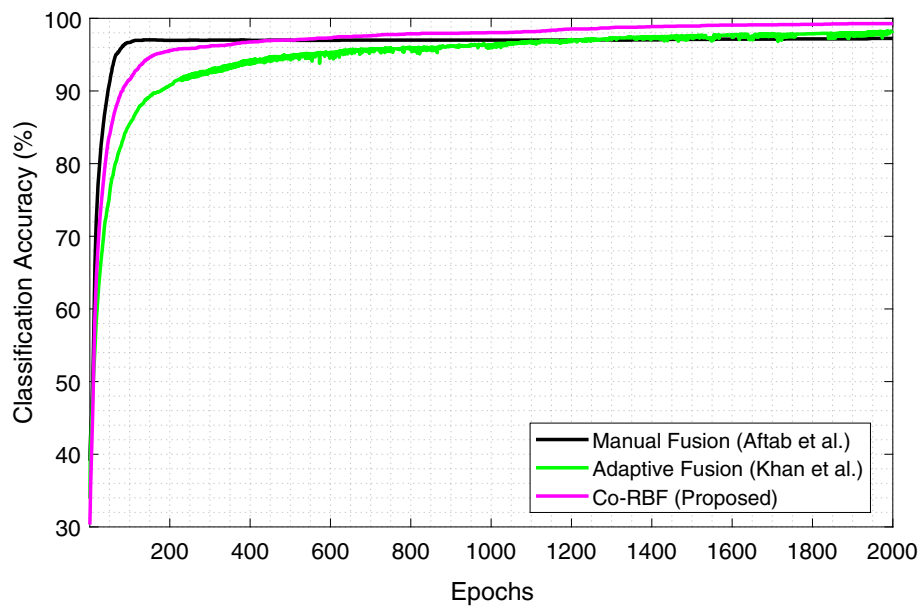


Fig. 4 Training accuracy curves of different RBF algorithms on Iris Flowers Dataset

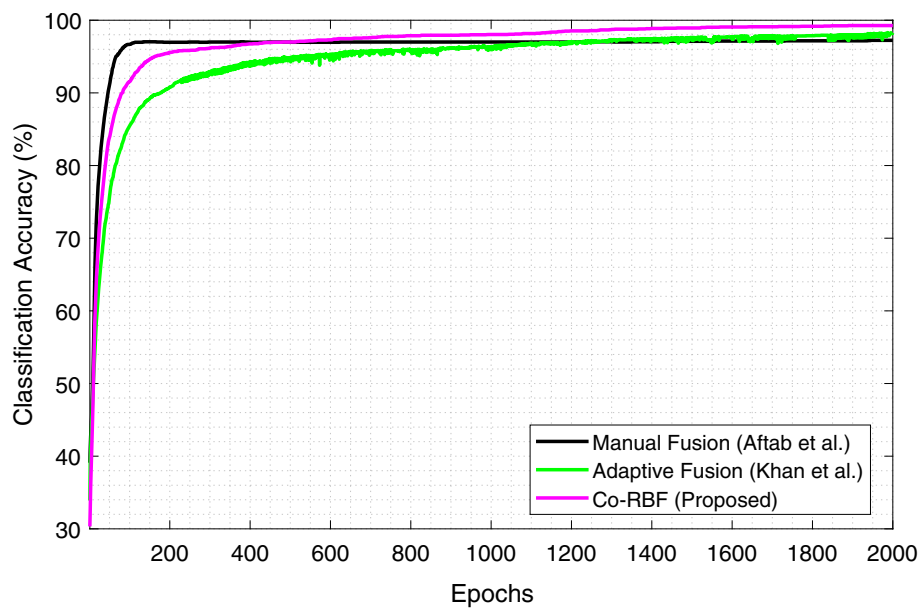


Fig. 5 Testing accuracy curves of different RBF algorithms on Iris Flowers dataset

Table 2 Average classification sensitivity (in %) of Iris Flowers obtained by different RBF algorithms after training for 2000 epochs

| Architecture | Phase | Virginica mean \pm std | Versicolor mean \pm std | Setosa mean \pm std |
|-------------------------------|----------|-----------------------------|------------------------------|--------------------------|
| Manual Fusion (Aftab et al.) | Training | 97.10 \pm 1.58 | 96.03 \pm 1.24 | 100 \pm 0.00 |
| | Testing | 100 \pm 0.00 | 100 \pm 0.00 | 91.00 \pm 3.02 |
| Adaptive Fusion (Khan et al.) | Training | 98.65 \pm 1.644 | 97.13 \pm 2.11 | 100 \pm 0.00 |
| | Testing | 100 \pm 0.00 | 97.40 \pm 13.83 | 98.10 \pm 3.94 |
| Co-RBF (Proposed) | Training | 97.55 \pm 0.35 | 97.50 \pm 0.00 | 100 \pm 0.00 |
| | Testing | 100 \pm 0.00 | 100 \pm 0.00 | 97.40 \pm 4.41 |

Table 3 Average classification specificity (in %) of Iris Flowers obtained by different RBF algorithms after training for 2000 epochs

| Architecture | Phase | Virginica mean \pm std | Versicolor mean \pm std | Setosa mean \pm std |
|-------------------------------|----------|-----------------------------|------------------------------|--------------------------|
| Manual Fusion (Aftab et al.) | Training | 98.01 \pm 0.62 | 98.55 \pm 0.79 | 100 \pm 0.00 |
| | Testing | 100 \pm 0.00 | 95.50 \pm 1.51 | 100 \pm 0.00 |
| Adaptive Fusion (Khan et al.) | Training | 98.56 \pm 1.06 | 99.33 \pm 0.82 | 100 \pm 0.00 |
| | Testing | 98.70 \pm 6.91 | 99.05 \pm 1.97 | 100 \pm 0.00 |
| Co-RBF (Proposed) | Training | 98.75 \pm 0.00 | 98.78 \pm 0.18 | 100 \pm 0.00 |
| | Testing | 100 \pm 0.00 | 98.70 \pm 2.20 | 100 \pm 0.00 |

Table 4 is showing Youden index of different algorithms on Iris Flowers dataset. It is a popular index used to quantified the overall capacity of the model for pattern classification. During the training phase, adaptive kernel fusion [25] attained the best indices of 0.9721, 0.9646 and 1.0000 for Virginica, Versicolor and Setosa classes respectively. Followed by our algorithm with indices of 0.9630 (0.0091 less than the best), 0.9628 (0.0018 less than the best) and 1.0000 for Virginica, Versicolor and Setosa classes respectively. Manual kernel fusion [1] is in the last with indices of 0.9511, 0.9458 and 1.0000 for Virginica, Versicolor and Setosa classes respectively.

During testing phase, our algorithm achieved the best Youden indices of 1.0000 and 0.9870 for classes Virginica and Versicolor respectively. However, it attained the second best Youden index of 0.9740 on Setosa class (i.e. 0.0070 less than 0.9810 the best Youden index reached by adaptive kernel fusion [25]). In the light of our simulation results of Virginica and Versicolor classes, adaptive kernel fusion [25] is the second best (with Youden indices of 0.9870 and 0.9745 for Virginica and Versicolor classes respectively) and manual kernel fusion [1] is the worst (with Youden indices of 1.0000 and 0.9550 for Virginica and Versicolor classes respectively) in term of Youden index during testing phase.

4.2 Function Approximation Problem

Function approximation is a way to describe the behavior of complicated functions using available observations from the domain through ensembles of simpler functions. It has special importance in several research domains like dynamic system modeling, nonlinear complex-valued signal processing, and biological activity modeling etc [22, 38, 47].

Table 4 Average Youden index of Iris Flowers obtained by different RBF algorithms after training for 2000 epochs

| Architecture | Phase | Virginica | Versicolor | Setosa |
|-------------------------------|----------|---------------|---------------|---------------|
| Manual Fusion (Aftab et al.) | Training | 0.9511 | 0.9458 | 1.0000 |
| | Testing | 1.0000 | 0.9550 | 0.9100 |
| Adaptive Fusion (Khan et al.) | Training | 0.9721 | 0.9646 | 1.0000 |
| | Testing | 0.9870 | 0.9745 | 0.9810 |
| Co-RBF (Proposed) | Training | 0.9630 | 0.9628 | 1.0000 |
| | Testing | 1.0000 | 0.9870 | 0.9740 |

For the function approximation problem, we consider the following non linear function defined as:

$$f(x_1, x_2) = e^{(x_1^2 - x_2^2)}, \quad \forall -1 \leq x_1 \leq 1 \text{ and } -1 \leq x_2 \leq 1, \quad (58)$$

For training phase, x_1 and x_2 were selected over the interval $[-1, 1]$ with sampling spacing of 0.2. Whereas for the testing phase, x_1 and x_2 were selected over the interval $[-0.9, 0.9]$ at the same rate. Hence, 121 and 100 samples were used for training and testing respectively.

All the RBF algorithms were initialized with the following specifications. Learning rate was set to 1×10^{-3} and the Gaussian kernel spread was taken to be unity. All 121 hidden neurons were configured by selecting training samples as centers for the kernel. Weights and bias were initialized randomly for every run.

MSE curves of different RBF algorithms during training are shown in Fig. 6. Adaptive kernel fusion architecture [25] showed the highest convergence rate for first 50 epochs but then got stuck in a local minima and achieved the higher error of -20.5 db at 2000 epochs. In contrast, our proposed architecture showed moderate but consistent convergence rate thus achieved the minimum error -39.83 dB at 2000 epochs. Moreover, manual kernel fusion architecture [1] exhibited moderate final convergence by attaining the error of -36.53 db at 2000 epochs.

Instantaneous error of our proposed architecture is well bounded between -0.1 and 0.1 whereas that of manual kernel fusion [1] is bounded between -0.15 and 0.15 and that of Adaptive kernel fusion [25] is bounded between 4.5 and -3.0 as depicted in 8. Hence, Adaptive kernel fusion [25] is the worst in term of instantaneous error among all the three algorithms. As the result, the predicted output of our proposed architecture mapped the actual output in the best manner as showed in Fig. 7.

Figures 9 and 10 are showing the error surfaces of different RBF algorithms on training and testing data. Error surface of Adaptive kernel fusion [25] is quite spiky for both the training and testing data i.e. bounded between 4.5 and -3.0 (training data) and 8.0 and -3.5 (testing data) respectively. It indicates that the algorithm poorly approximated the given function. In contrast, error surfaces of our proposed architecture are very flat bounded between 1.0 and -1.0 in case training data and that -0.12 and -0.14 in case of testing data. This indicates that given function is well approximated by Co-RBFNN. Manual kernel fusion is moderately spiky with error bound of $(-0.15, 0.15)$ for training data and that of $(-0.22, 0.13)$ for testing data. Thus, its ability of function approximation of the given function is average.

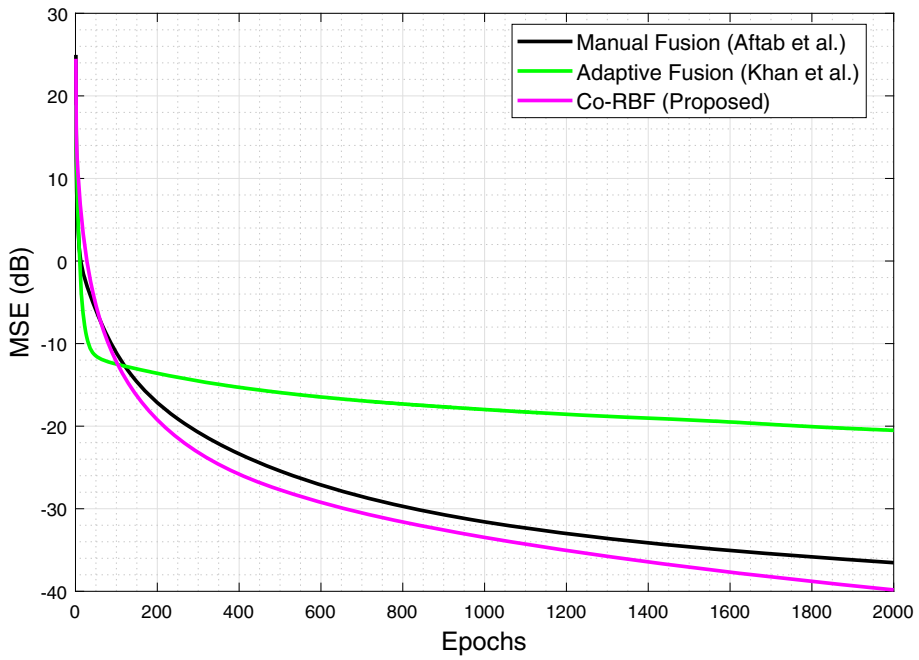


Fig. 6 MSE curves of different RBF algorithms on function approximation problem

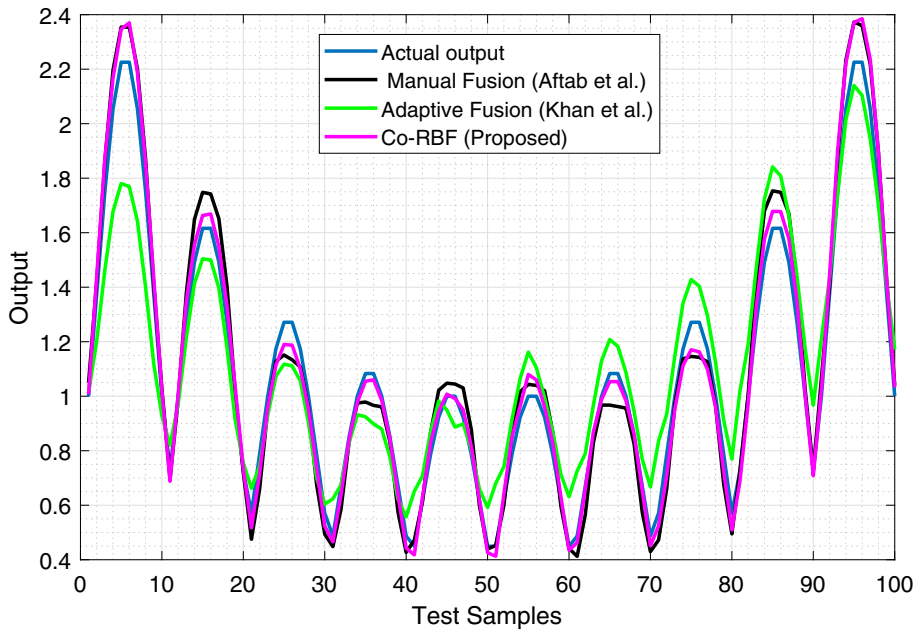


Fig. 7 Predicted output of different RBF algorithms on test data of function approximation problem

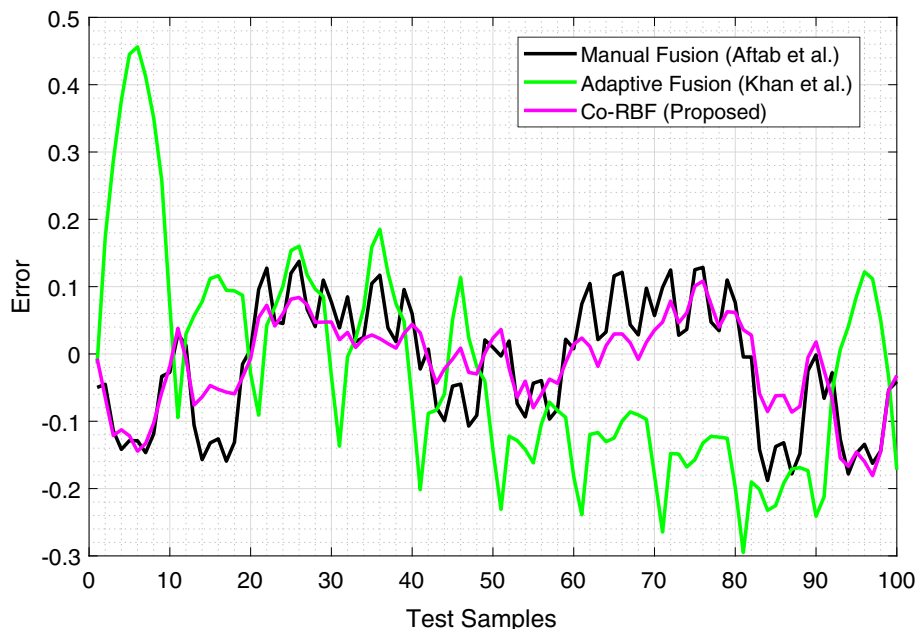


Fig. 8 Instantaneous error of different RBF algorithms on test Data of function approximation problem

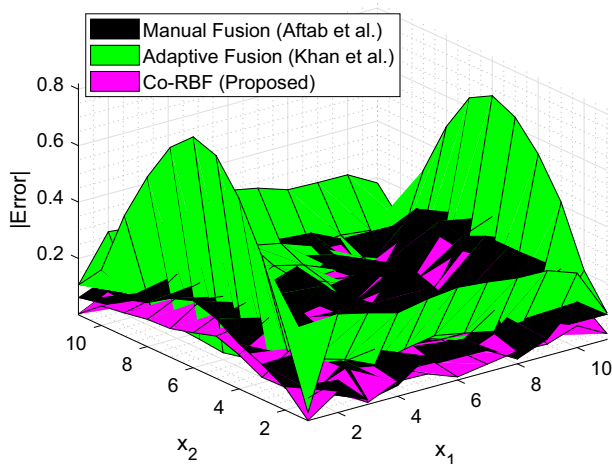


Fig. 9 Error surfaces of different RBF algorithms on train data of function approximation Problem

4.3 Nonlinear System Identification

System identification/nonlinear system identification is a systematic approach to build mathematical models of dynamic systems using measurements of only the system's input and output signals. It has several applications in diverse fields ranging from wireless communication systems [2, 21, 37] to geo localization of mines [32] etc. It is considered to be a highly challenging research problem in the domain of signal processing and can be effectively

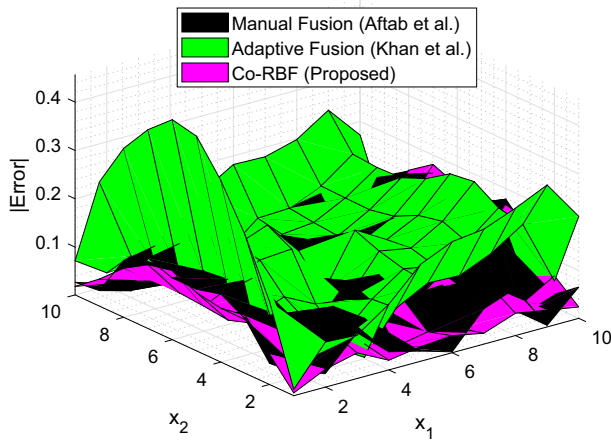


Fig. 10 Error surfaces of different RBF algorithms on test data of function approximation problem

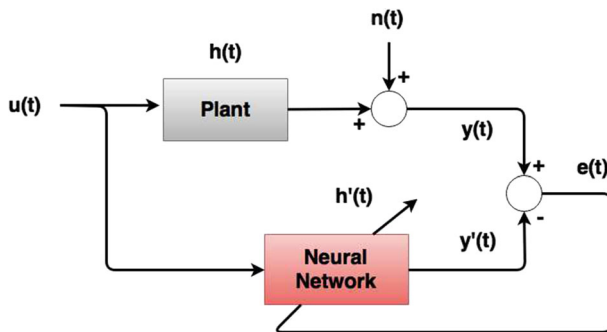


Fig. 11 Nonlinear system identification using RBF neural network

addressed using neural networks [19]. Fig. 11 depicts a general systematic approach used by the RBF neural networks for this purpose. For the evaluation of the proposed architecture, we consider a first order non linear system defined by the following equation:

$$y_t = 2u_{(t)} - 0.5u_{(t-1)} - 0.1u_{(t-2)} - 0.7(\cos(3u_{(t)})) + e^{-|u_{(t)}|}, \quad (59)$$

where, u_t and y_t are the system input and output respectively. The input signal is a unit amplitude square wave of length 400 samples and 50% duty cycle. For model estimation, during training phase a Gaussian noise of zero mean and 0.2 variance was added.

The following specifications are used for the RBF algorithms: (1) a learning rate of 1×10^{-4} , (2) the Gaussian kernel spread is set to 0.5, and (3) for 5 neurons, the centers are selected as $\mathbf{m} = \{-100, -50, 0, 50, -100\}$.

MSE curves of different RBF algorithms are depicted in Fig. 12. The proposed architecture yields the highest convergence rate with a minimum error of 3.48 dB which is identical to the manual and adaptive fusion method [1, 25]. Comparison of the actual and estimated test signals for the different RBF algorithms is illustrated in Fig. 13. In an inset plot, it is evident that our proposed algorithm estimates the actual test signal significantly better compared to the other algorithms.

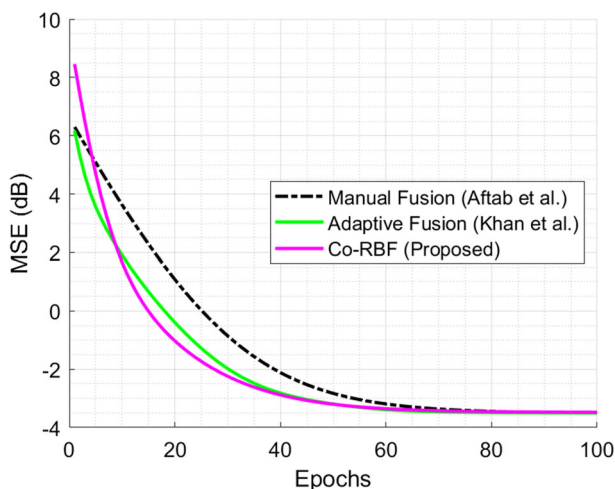


Fig. 12 MSE curves of different RBF algorithms on system identification problem

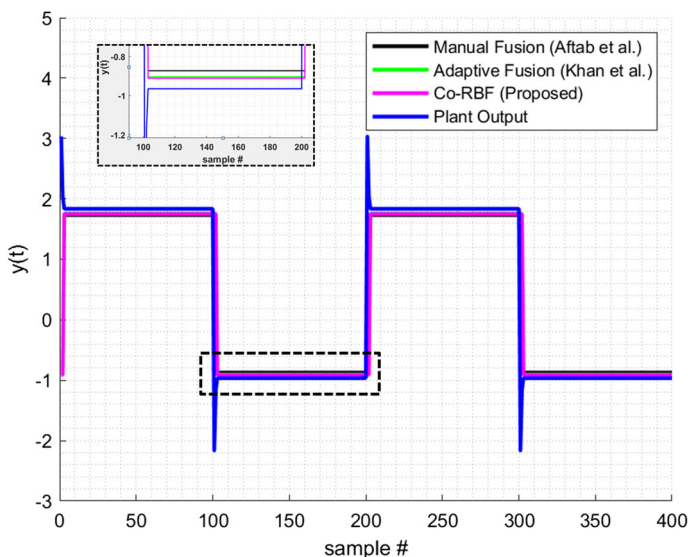


Fig. 13 Estimated output of different the RBF algorithms on test data of system identification problem

5 Conclusion

In this paper, we proposed a novel multi-kernel RBF neural network architecture called Co-RBFNN. The proposed kernel fusion method uses matrix-based mixing weights enabling each (primary and sub-primary) kernel to learn independent weights. A graphical explanation highlighting the underlying reasons for the improvement is provided along with a detailed mathematical analysis. We demonstrated the efficacy of the proposed solution on three important problems, namely: (i) Nonlinear system identification, (ii) pattern classification and (iii) function approximation. The proposed algorithm has shown to comprehensively

outperform the two state-of-the-art methods i.e. manual and adaptive fusion of kernels. For the problem of pattern classification, the proposed framework achieved the lowest error floor of -35.39 dB after 2000 epochs of training. For the testing phase the proposed Co-RBFNN achieved a high classification accuracy of 99.13% (approximately) which compares favorably with the contemporary methods. For the function approximation problem, our proposed method converged to the lowest error of -39.83 dB after 2000 epochs. The convergence rate of the proposed algorithm was also found to be better than the competing methods. For the nonlinear system identification problem, the proposed Co-RBFNN algorithm exhibited the fastest convergence rate achieving a minimum error of -3.48 dB. The unseen test signal was more accurately estimated by the proposed approach compared to the contemporary methods. MATLAB code for a sample problem can be downloaded from https://github.com/Shujaat123/Robust_RBF.

The proposed novel approach enables independent learning of the mixing weights making it superior compared to the contemporary approaches. However, one sophistication of the current method is that it requires fine-tuning and pre-processing of data, which requires some experience on behalf of inexperienced users. For such users, in future, we are interested in designing a toolbox version that can facilitate the adaptation of the proposed method. Additionally, it would be interesting to incorporate more sophisticated learning strategies such as evolutionary methods and expanding the domain of our experiments to other more practical problems.

Acknowledgements Syed Muhammad Atif acknowledges the support of HEC, Pakistan under Indigenous Ph.D. Fellowship Program (PIN 315-13358-2EG3-204).

Author Contributions All authors contributed to the study conception and design. Shujaat Khan designed the research, Syed Muhammad Atif and Shujaat Khan conducted and conceived the experiments and performed analysis, The initial draft of the manuscript was written by Syed Muhammad Atif and all authors commented on previous versions of the manuscript. All authors discussed the results and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions Open Access funding enabled and organized by CAUL and its Member Institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


1. Aftab W, Moinuddin M, Shaikh MS (2014) A Novel Kernel for RBF Based Neural Networks. Abstract and Applied Analysis 2014:1–10. <https://doi.org/10.1155/2014/176253>
2. Ahmad J, Khan S, Usman M, Naseem I, Moinuddin M, Syed HJ (2017) Fclms: Fractional complex lms algorithm for complex system identification. In: 2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), 39–43. IEEE
3. Alexandridis A, Chondrodima E, Sarimveis H (2016) Cooperative learning for radial basis function networks using particle swarm optimization. Applied Soft Computing 49:485–497. <https://doi.org/10.1016/j.asoc.2016.08.032>

4. Aljarah I, Faris H, Mirjalili S, Al-Madi N (2018) Training radial basis function networks using biogeography-based optimizer. *Neural Computing and Applications* 29(7):529–553. <https://doi.org/10.1007/s00521-016-2559-2>
5. de Almeida Rego JB, de Medeiros Martins A, Costa EdB (2014) Deterministic System Identification Using RBF Networks. *Mathematical Problems in Engineering* 2014:1–10. <https://doi.org/10.1155/2014/432593>
6. Bu K, He Y, Jing X, Han J (2020) Adversarial transfer learning for deep learning based automatic modulation classification. *IEEE Signal Processing Letters*
7. Bucak SS, Jin R, Jain AK (2014) Multiple Kernel Learning for Visual Object Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7):1354–1369. <https://doi.org/10.1109/TPAMI.2013.212>
8. Chen ZY, Kuo RJ (2019) Combining SOM and evolutionary computation algorithms for RBF neural network training. *Journal of Intelligent Manufacturing* 30(3):1137–1154. <https://doi.org/10.1007/s10845-017-1313-7>
9. Fisher RA (1936) THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics* 7(2):179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
10. Fu L, Zhang M, Li H (2010) Sparse RBF Networks with Multi-kernels. *Neural Processing Letters* 32(3):235–247. <https://doi.org/10.1007/s11063-010-9153-x>
11. Gan M, Peng H, Dong Xp (2012) A hybrid algorithm to optimize RBF network architecture and parameters for nonlinear time series prediction. *Applied Mathematical Modelling* 36(7):2911–2919. <https://doi.org/10.1016/j.apm.2011.09.066>
12. Gao F, Han L (2012) Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications* 51(1):259–277. <https://doi.org/10.1007/s10589-010-9329-3>
13. Goodfellow I (2016) Nips 2016 tutorial: Generative adversarial networks. arXiv preprint [arXiv:1701.00160](https://arxiv.org/abs/1701.00160)
14. Gu Y, Liu T, Jia X, Benediktsson JA, Chanussot J (2016) Nonlinear Multiple Kernel Learning With Multiple-Structure-Element Extended Morphological Profiles for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* 54(6):3235–3247. <https://doi.org/10.1109/TGRS.2015.2514161>
15. Hassan AK, Moinuddin M, Al-Saggaf UM, Shaikh MS (2018) On the Kernel Optimization of Radial Basis Function Using Nelder Mead Simplex. *Arabian Journal for Science and Engineering* 43(6):2805–2816. <https://doi.org/10.1007/s13369-017-2888-1>
16. Haykin SS (1999) *Neural networks: a comprehensive foundation*, 2nd edn. Prentice Hall, Upper Saddle River, N.J
17. Haykin SS (2014) *Adaptive filter theory*, 5th edn. Pearson, Upper Saddle River, New Jersey
18. Ibrahim MS, Dong W, Yang Q (2020) Machine learning driven smart electric power systems: Current trends and new perspectives. *Applied Energy* 272:115,237
19. Khan S, Ahmad J, Naseem I, Moinuddin M (2018) A novel fractional gradient-based learning algorithm for recurrent neural networks. *Circuits, Systems, and Signal Processing* 37(2):593–612
20. Khan S, Ahmad J, Sadiq A, Naseem I, Moinuddin M (2018) Spatio-Temporal RBF Neural Networks. In: 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), pp. 1–5. IEEE, Karachi, Pakistan. <https://doi.org/10.1109/ICEEST.2018.8643322>
21. Khan S, Ahmed N, Malik MA, Naseem I, Togneri R, Bennamoun M (2017) Flmf: Fractional least mean fourth algorithm for channel estimation in non-gaussian environment. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC), 466–470. IEEE
22. Khan S, Huh J, Ye JC (2019) Universal plane-wave compounding for high quality us imaging using deep learning. In: 2019 IEEE International Ultrasonics Symposium (IUS), 2345–2347. IEEE
23. Khan S, Huh J, Ye JC (2020) Adaptive and compressive beamforming using deep learning for medical ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 1–1
24. Khan S, Naseem I, Malik MA, Togneri R, Bennamoun M (2018) A fractional gradient descent-based rbf neural network. *Circuits, Systems, and Signal Processing* 37(12):5311–5332
25. Khan S, Naseem I, Togneri R, Bennamoun M (2017) A novel adaptive kernel for the rbf neural networks. *Circuits, Systems, and Signal Processing* 36(4):1639–1653
26. Khan S, Naseem I, Togneri R, Bennamoun M (2018) Rafp-pred: Robust prediction of antifreeze proteins using localized analysis of n-peptide compositions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15(1):244–250
27. Lee D, Kim J, Moon WJ, Ye JC (2019) Collagan: Collaborative gan for missing image data imputation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2487–2496

28. Liu Y, Zhao J, Xiao Y (2018) C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network. *Neurocomputing* 275:733–746. <https://doi.org/10.1016/j.neucom.2017.09.015>
29. Meng X, Rozycki P, Qiao JF, Wilamowski BM (2018) Nonlinear System Modeling Using RBF Networks for Industrial Application. *IEEE Transactions on Industrial Informatics* 14(3):931–940. <https://doi.org/10.1109/TII.2017.2734686>
30. Muhammad M, Naseem I, Aftab W, A Bencherif S, Memich A (2017) A Weighted Cosine RBF Neural Networks. *J Mol Biol Biotech* 2(2): 1–8. URL <http://www.imedpub.com/articles/a-weighted-cosine-rbf-neural-networks.pdf>
31. Naseem I, Khan S, Togneri R, Bennamoun M (2017) Ecmsrc: A sparse learning approach for the prediction of extracellular matrix proteins. *Current Bioinformatics* 12(4):361–368
32. Nerguizian C, Despins C, Affès S (2006) Geolocation in mines with an impulse response fingerprinting technique and neural networks. *IEEE transactions on wireless communications* 5(3):603–611
33. Pal NR, Chakraborty D (2000) Mountain and subtractive clustering method: Improvements and generalizations. *International Journal of Intelligent Systems* 15(4):329–341. 10.1002/(SICI)1098-111X(200004)15:4<329::AID-INT5>3.0.CO;2-9. URL <http://doi.wiley.com/10.1002/%28SICI%291098-111X%28200004%2915%3A4%3C329%3A%3AAID-INT5%3E3.0.CO%3B2-9>
34. Peng S, Jiang H, Wang H, Alwageed H, Zhou Y, Sebdani MM, Yao YD (2018) Modulation classification based on signal constellation diagrams and deep learning. *IEEE transactions on neural networks and learning systems* 30(3):718–727
35. Pratiwi M, Alexander, Harefa J, Nanda S (2015) Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network. *Procedia Computer Science* 59: 83–91. <https://doi.org/10.1016/j.procs.2015.07.340>. URL <https://linkinghub.elsevier.com/retrieve/pii/S1877050915018694>
36. Sadiq A, Ibrahim MS, Usman M, Zubair M, Khan S (2018) Chaotic time series prediction using spatio-temporal rbf neural networks. In: 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 1–5. IEEE
37. Sadiq A, Khan S, Naseem I, Togneri R, Bennamoun M (2019) Enhanced q-least mean square. *Circuits, Systems, and Signal Processing* 38(10):4817–4839
38. Sikora R, Giza Z, Filipowicz F, Sikora J (2000) The bell function approximation of material coefficients distribution in the electrical impedance tomography. *IEEE Transactions on Magnetics* 36(4):1023–1026
39. Simon D (2008) Biogeography-Based Optimization. *IEEE Transactions on Evolutionary Computation* 12(6):702–713. <https://doi.org/10.1109/TEVC.2008.919004>
40. Tuia D, Camps-Valls G, Matasci G, Kanevski M (2010) Learning Relevant Image Features With Multiple-Kernel Classification. *IEEE Transactions on Geoscience and Remote Sensing* 48(10):3780–3791. <https://doi.org/10.1109/TGRS.2010.2049496>
41. Usman M, Khan S, Lee JA (2020) Afp-lse: Antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs. *Scientific Reports* 10(1):1–13
42. Varma M, Babu BR (2009) More generality in efficient multiple kernel learning. In: Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09, pp. 1–8. ACM Press, Montreal, Quebec, Canada. <https://doi.org/10.1145/1553374.1553510>. URL <http://portal.acm.org/citation.cfm?doid=1553374.1553510>
43. Vetrivel A, Gerke M, Kerle N, Nex F, Vosselman G (2018) Disaster damage detection through synergistic use of deep learning and 3d point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 140:45–59. <https://doi.org/10.1016/j.isprsjprs.2017.03.001>
44. Wen Z, Xie L, Feng H, Tan Y (2019) Robust fusion algorithm based on RBF neural network with TS fuzzy model and its application to infrared flame detection problem. *Applied Soft Computing* 76:251–264. <https://doi.org/10.1016/j.asoc.2018.12.019>
45. Yang X, Li Y, Sun Y, Long T, Sarkar TK (2018) Fast and Robust RBF Neural Network Based on Global K-means Clustering with Adaptive Selection Radius for Sound Source Angle Estimation. *IEEE Transactions on Antennas and Propagation* 1–1. <https://doi.org/10.1109/TAP.2018.2823713>. URL <http://ieeexplore.ieee.org/document/8335765/>
46. Yang XS (2010) Nature-inspired metaheuristic algorithms, 2nd edn. Luniver Press, Frome
47. Yoon YH, Khan S, Huh J, Ye JC (2018) Efficient b-mode ultrasound image reconstruction from sub-sampled rf data using deep learning. *IEEE transactions on medical imaging* 38(2):325–336
48. Zhu JZ, Cao JX, Zhu Y (2014) Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections. *Transportation Research Part C: Emerging Technologies* 47:139–154. <https://doi.org/10.1016/j.trc.2014.06.011>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Syed Muhammad Atif¹ · Shujaat Khan²  · Imran Naseem^{3,4}  · Roberto Togneri⁴ · Mohammed Bennamoun⁵

Syed Muhammad Atif
s.m.atif@pafkiet.edu.pk

Shujaat Khan
shujaat@kaist.ac.kr

Roberto Togneri
roberto.togneri@uwa.edu.au

Mohammed Bennamoun
mohammed.bennamoun@uwa.edu.au

- ¹ Graduate School of Science and Engineering, Karachi Institute of Economics and Technology, Korangi Creek, Karachi 75190, Pakistan
- ² Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
- ³ College of Engineering, Karachi Institute of Economics and Technology, Korangi Creek, Karachi 75190, Pakistan
- ⁴ School of Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia
- ⁵ School of Physics, Mathematics and Computing, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia