Proposal of SVM Utility Kernel for Breast Cancer Survival Estimation

Nikhilanand Arya, Archana Mathur, Snehanshu Saha, and Sriparna Saha

Abstract—The advancement of medical research in the field of cancer prognosis and diagnosis using various modalities has put oncologists under tremendous stress. The complexity and heterogeneity involved in multiple modalities and their significantly varied clinical outcomes make it difficult to analyze the disease and provide the correct treatment. Breast cancer is the major concern among all cancers worldwide, specifically for females. To help oncologists and cancer patients, research for breast cancer survival estimation has been proposed. It ranges from complex deep neural networks to simple and interpretable architectures. We propose a utility kernel for a support vector machine (SVM) in this article. It is a simple yet powerful function, which performs better than other popular machine learning algorithms and deep neural networks in the task of breast cancer survival prediction using the TCGA-BRCA dataset. This study validates the proposed utility kernel using four different modalities (gene expression, copy number variation, clinical, and histopathological tissue images) and their multi-modal combinations. The SVM based on our utility kernel empirically proves its efficacy by achieving the highest value on various performance measures, whereas advanced deep neural networks fail to train on small and highly imbalanced breast cancer data.

Index Terms—Breast cancer survival estimation, gene expression, copy number variation, histopathological whole slide images, utility kernel, support vector machine, machine learning, deep neural networks

1 Introduction

Breast is the hub of different types of tissues like fatty tissue and dense tissue. Each tissue is the network of lobes, lobules, and milk glands. Blood and lymph vessels connected with lymph nodes are parts of the breast that nourish breast cells and manage bodily waste products. Breast cancer originates once breast cells start growing uncontrollably, resulting in tumor formation. The spread of breast cancer can be life-threatening if it becomes metastasis breast cancer. It starts spreading into adjacent organs through blood or lymph vessels. Biologically, breast cancer are of two types, invasive and non-invasive. Breast cancer that spreads to nearby tissues and/or distant organs belongs to invasive breast cancer. Non-invasive breast cancer does not spread beyond the milk ducts or lobules of the breast. The Global Cancer Observatory (GCO)[1] "an interactive web-based platform presenting global cancer statistics to inform cancer control and research" under the World Health Organization (WHO) has estimated 19,292,789 new cancer cases worldwide, both sexes, all ages in the year 2020. Breast cancer alone has 11.7% patients out of the estimated cancer patients. Once we narrowed our search by filtering only female patients, breast cancer cases increased to 24.5% of all female cancer estimated cases in 2020. The expected number of breast cancer cases by 2040 in females is 3.2 million, which is currently 2.3 million, and the expected number of breast cancer deaths is 1 million. Considering the alarming future of breast cancer, we need to have a better prognosis and diagnosis tool which uses both qualitative information from histology and quantitative information from genomic data to predict the clinical outcomes. A correct and timely prognosis of cancer patients can assist both patients and oncologists in choosing appropriate treatment. The appropriate treatment helps in avoiding the toxic side effects of various cancer-related therapies, preventing over-treatment, thereby reducing economic costs [1], more effectively including and excluding patients in a randomized trial, and developing palliative care and hospice care systems.

As a result, predicting survival has emerged as a prominent concern in modern breast cancer research. The complexities and significantly varied clinical outcomes due to histology and genomic data make it difficult to predict and treat [2]. Recent advances in medical imaging and next-generation sequencing technologies help the oncologist to rapidly and extensively evaluate a large number of genes and samples to predict breast cancer's diagnosis, prognosis, and potential response to therapy. Nowadays, METABRIC [3], The Cancer Genome Atlas (TCGA) [4] etc. are making a large number of genome-scale transcriptomic data publicly available for breast cancer, which is being used for the

1. https://gco.iarc.fr/

- *Nikhilanand Arya and Sriparna Saha are with the Department of Computer Science & Engineering, Indian Institute of Technology Patna, Patna, Bihar 801106, India. E-mail: nikhilaryan92@gmail.com, sriparna@iitp.ac.in.*
- *Archana Mathur is with the Department of Information Science and Engineering, Nitte Meenkashi Institute of Technology and Management, Bangalore, Karnataka 560064, India. E-mail: mathurarchana77@gmail.com.*
- *Snehanshu Saha is with the Department of Computer Science & Information Systems and APPCAIR, BITS Pilani, Sancoale, Goa 403726, India. E-mail: snehanshu.saha@ieee.org.*

Manuscript received 1 November 2021; revised 20 June 2022; accepted 26 July 2022. Date of publication 22 August 2022; date of current version 3 April 2023.
The work of Snehanshu Saha was supported by the DBT-Builder project, Govt. of India under Grant BT/INF/22/SP42543/2021, and in part by SERB-EMR under Grant EMR/05687.
(Corresponding author: Sriparna Saha.)
Digital Object Identifier no. 10.1109/TCBB.2022.3198879

development and validation of various computational methods related to breast cancer survival analysis. Survival is defined as the amount of time a patient lives after being diagnosed with an illness. The 5-year criterion is critical for standardizing reporting and determining survivability. Because it takes at least 5 years to label a patient record as survived or not survived, some earlier research utilized a 5-year criterion to determine the cohort's survivability [5]. Breast cancer is a complex disease, and while survival rates have significantly grown in recent years, the 5-year survival rate varies greatly between individuals [6]. In this study, we have used varying survival years in the range of 5 years to 9 years as survival cut-offs, and patients having life expectancy less than survival cut-off and more than survival cut-off are classified as short-term survivors and long-term survivors, respectively. Numerous advanced and unique approaches for the application of big data analysis techniques in the creation of survival prediction models have been developed as medical research has progressed. Machine learning (ML) has the ability to automatically learn data models, requires no implicit assumptions, and can handle dependency and nonlinear interactions between variables [7]. It excels at dealing with a huge number of complex higher-order interactions found in medical data. As a result, machine learning techniques have a lot of potential for use in everyday medical practice as leading health informatics tools.

2 Literature Survey

Past few years have seen plenty of research on the prognosis and diagnosis of breast cancer. These researches were initially motivated by the significance of microarray technology to understand the gene expression profile of breast cancer patients, however, only a small fraction shows clear prognostic significance [8] [9]. The identification task of 70-gene markers which are responsible for breast cancer is performed by *Van't Veer* et al. [8]. They used microarray data with primary breast tumors from 117 patients and utilized a supervised classification. In the extension of gene markers identification, *Van* et al. [10] validated the efficacy of already proposed 70-gene prognostic signature over 295 breast cancer patients. Some uni-modal machine learning classification methods, such as Support Vector Machine [11], Bayes classifier [12] and Random Forest (RF) [13] have been used to predict breast cancer prognosis prediction and survival estimation. For instance, *Xu* et al. [11] proposed a SVM based recursive feature elimination technique to identify 50 gene-markers of breast cancer prognosis prediction. *Nguyen* et al. [13] proposed the use of tissue images in breast cancer diagnosis and prognosis, which outperforms previously reported results. They used Bayesian probability to rank the features extracted from tissue images and random forests for the final prognosis prediction task. All these techniques were using a single modality either gene expression or tissue images for the prediction. With the availability of multi-modal data on cancer patients, the research focus has been shifted from uni-modal to multi-modal. In the early days, some researchers used gene-expression and clinical details as two different modalities and developed bi-modal architectures. *Sun* et al. proposed I-RELIEF [14] feature selection algorithm to select hybrid markers consisting of three genes and two clinical markers for the breast cancer prognosis. Further, *Gevaert* et al. [12] designed a bi-modal probabilistic architecture, which uses the power of the Bayesian Network to prognosticate lymph-node negative breast cancer using pathological and gene expression data. A thorough literature review highlighted that gene expression profile is conducive to breast cancer prognosis prediction, but high dimensional (approx 25,000 genes per patient) microarray data and correlation between the genes puts some serious challenges in the gene-markers identification and related prognosis prediction. To reduce the dimensions of microarray data and prognosticate breast cancer with improved accuracy, a probabilistic graphical model (PGM) was proposed by *Khademi* et al. [15]. They used the popular dimensionality reduction technique, PCA followed by the deep belief network to extract feature representation from microarray data. It is a bi-modal study with the combination of two independent models for microarray and clinical profiles.

With the evolution of medical imaging technology [16], deciphering abnormal pictures and understanding tumor morphology has never been easy. It results in the incorporation of more cancer-related modalities (images) to improve breast cancer survival prediction performance. In recent research, few computational methods for predicting cancer clinical outcomes based on pathological pictures were developed, assuming that pathological images may give additional information about tumor features. Using indicative markers from histopathology slides, *Wang* et al. [17] provided an integrated framework for non-small cell lung cancer computer-aided diagnosis and survival analysis. For lung cancer survival prediction, *Zhu* et al. [18] created a prediction model that combined pathological image characteristics with gene expression signatures. *Yu* et al. [19] extracted 9879 meaningful image characteristics from 2186 Hematoxylin and Eosin pathological whole-slide images (WSIs) of non-small cell lung cancer and used conventional classification methods to separate short-term and longer-term survivors.

Despite the positive results of the above-mentioned methodologies for lung cancer, there is currently a dearth of studies with pathological imaging for breast cancer clinical outcome analysis due to the complexity and heterogeneity of this serious disease. Meanwhile, the constantly growing number of characteristics from various data sources and usage of heterogeneous features may provide a significant difficulty in combining them efficiently for breast cancer survival prediction. Recent advances in deep learning algorithms have demonstrated that a model with many modalities of input data source outperforms a model with a single source of input data. This finding has been supported by multi-modal architectures in various investigations like diseased gene prognosis by DeePROG[20], protein function prediction by MultiPredGO [21] and breast cancer prognosis prediction by MDNNMD [22], STACKED RF [23] and SiGaAtCNN + STACKED RF [24]. *Sun* et al. also created the GPMKL [25] approach based on the ensemble of multiple kernel functions, which integrates genomic data and pathological images to predict breast cancer prognosis. Considering the requirement for the study of survival analysis in the medical sphere and the availability of histopathology images, *Tang* et al. have built a Capsule Network called

CapSurv [26]. This method uses unique loss function called survival loss, which works specifically for cancer patient survival analysis. The incomplete multi-view is a common issue while dealing with multi-modal data. Considering this limitation, *Arya and Saha* proposed a generative multi-modal architecture [27] for the breast cancer prognosis prediction task.

3 Motivation and Contribution

All the previous studies are based either on traditional machine learning methods, or on advanced deep learning techniques, or an ensemble of deep learning with machine learning. It is conspicuous from the literature review that ML was unable to improve the survival prediction beyond a particular limit due to the complexity and heterogeneity involved in multi-modal data. Deep neural networks came as a savior and further improved the performance of breast cancer survival prediction tasks. Few researchers also proposed the ensemble of deep learning with machine learning to have more robust and powerful architectures. In the race to have a more accurate and compelling model, we left our goal of designing an interpretable architecture, which is the main characteristic of classical ML algorithms. Nowadays, advanced deep neural architectures are complex enough to overfit small-sized data. They are unable to handle high-class imbalance. Preliminary investigation using several deep network architectures including Stacked and Variational Auto-Encoders (VAE) reveals that these DL models failed to learn anything over the WSI data modality. The AUC-ROC value was coming out to be 50% to 55%, a statistic slightly better than 'coin toss'. Considering these (anticipated) drawbacks and limitations of deep learning algorithms, we developed the SVM Utility Kernel. In this study, the proposed utility kernel is employed for the task of breast cancer survival prediction at various survival cut-offs (5 years - 9 years) and all possible fifteen combinations of clinical (CLN), gene expression (EXP), copy number variation (CNV), and histopathological whole slide tissue images (WSI) input modalities. The proposed kernel not only outperforms popular SVM kernels but also proves its efficacy in small-sized high class imbalanced breast cancer survival prediction. It shows drastic improvement in breast cancer survival prediction using histopathological WSIs, where various deep neural networks and convolutional neural networks unable to learn and predict anything. Additionally, the utility kernel satisfies the Mercer conditions, and the bounds on the generalization error (a measure of how the kernel performs on the unseen data samples) and its VC dimensions are thoroughly investigated in the current study.

4 Theoretical Framework

Introduction to Support Vector Machine (SVM)

Let $x \in R^n$ be linearly separable training samples in n dimensional space and $y \in \{-1, +1\}$. The hypothesis behind the SVM formulation is a hyperplane, characterized by w, vector perpendicular to the hyperplane, and b, that separates all the R^n into two halves. There could be many such hyperplanes, where the optimal hyperplane is the one with the largest

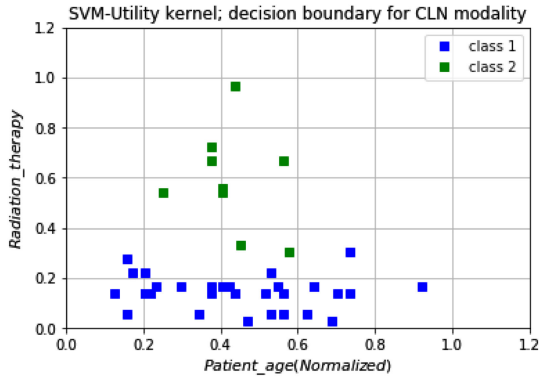

Fig. 1. Scatter plot of data taken from CLN modality for two features.

distance between the separating points of two classes (margin, m). The optimization problem is the following:

$$\underset{w}{minimize} \frac{1}{2}\|w\|^2$$

$$\text{subject to} \quad (< x_i, w > +b) \cdot y_i \geq 1 \quad \forall i = 1, ..., N$$

Assuming there are two hyperplanes (say, H_1 and H_2) that separate the data points with the largest margin m, then the optimization problem of maximizing m is solved by using the method of Lagrange multiplier. Furthermore, the problem reconstructs in its dual form as, (note: the concept of duality principle for the current Lagrangian equation entails the optimization problem perceived in two forms: one is primal and the other is it's dual.)

$$\underset{\theta}{maximize} \sum_{i=1}^{m} \theta_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \theta_i \theta_j y_i y_j < x_i \cdot x_j >$$

$$\text{subject to} \quad \theta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} \theta_i y_i = 0$$

Hence, the minimization with respect to w, b, and maximization with respect to θ are primal and dual forms, respectively. Here, θ_i, θ_j are the Lagrange multipliers, y_i is the class label, and $< ., . >$ denotes the inner product of input samples, x_i, x_j.

4.1 Kernel Trick to Solve Problems for Non-Linear Data

The binary classification problem of breast cancer survival prediction turns out to be non-linearly separable (see Figs. 1 and 2). The native SVM unable to classify non-linearly separable data, for which the kernel trick is used, which maps the data points into higher dimensional space using a non-linear function. Once they are transformed into higher dimensional feature space, the features become linearly separable. Radial Basis Function (RBF), polynomial, and sigmoid are classical kernels. Recently, the combination of traditional kernels and hybrid kernels are used for different applications [28], [29], [30]. *Mehmet Gonen and Ethem Alpaydın* [31] used a similar approach, and *Sun Cuijuan* constructed a K-type kernel function [32]. However, these kernels lagged in robustness with regard to theoretical and analytical explanations. Analytically, for linearly separable

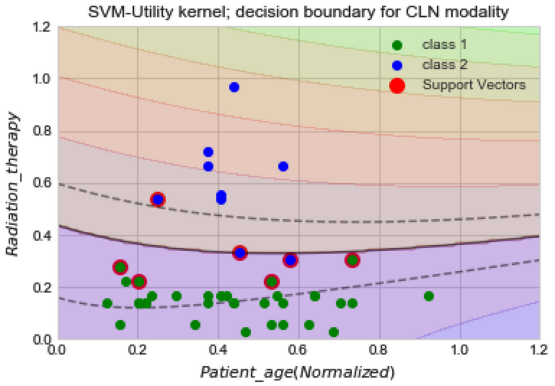

Fig. 2. Illustration of decision boundaries by applying utility kernel; $k_0 = 0.2, k_1 = 1.1, \alpha = 4$; Figure has seven support vectors around decision boundary for samples taken from CLN modality. class 1- long-term survivors, class -2 short-term survivors.

training data (x_j, y_j), for $j = 1....r$ we know that,

$$y_j = 1 \text{ if } x_j \in class1$$
$$y_j = -1 \text{ if } x_j \in class2$$

The optimal hyperplane is given as $h = w_1x_1 + ... + w_nx_n$ and the objective is to find $w_1, w_2....w_n$ with minimum $\|w\|^2$. Thus we embed $x_1, ...x_r$ in a much higher dimensional vector space, a Hilbert space which is a vector space endowed with an inner product allowing lengths and angles to be well defined (i.e., from $x = x_1, ..., x_{10}$ to $X = X_1, ..., X_{100000000}$). The embedding space has to be extremely large. Theoretically speaking, it's infinite-dimensional and therefore expensive to store computationally. It is important to note here that Hilbert space is infinite-dimensional.

Existing Kernels

A Radial basis kernel (RBF) non-linearly maps samples into a higher dimensional space. Values for the kernel K lie between 0 and 1 in contrast to polynomial kernels of which kernel values may go to zero or infinity [33]. The linear kernel is preferred when the number of features is large. The polynomial kernel projects the similarity of input vectors in a feature space over polynomials of varying degrees of the original variables. Polynomial kernel suffers from numerical instability problem. When $x^Ty + c \leq 1$ then $K(x_i, x_j)$ tends to zero and similarly when $x^Ty + c \geq 1$ then $K(x_i, x_j)$ tends to infinity. We seek to address this problem in the novel utility kernel described in the following subsection.

4.2 A Novel Kernel Function - Utility Kernel

Motivated from the field of Economics, we propose a novel kernel with the name Utility Kernel, expressed as:

$$K(x_i, x_j) = k_0 + k_1 < x_i, x_j >^{\alpha}$$

where k_0, k_1 and α are kernel parameters. An euclidean inner product (extensible to Hilbert Space) in $\Re^2$ with two arbitrary vectors x, y is defined as $\langle x, y \rangle = x_1y_1 + x_2y_2$. Given these characteristics, a class of utility functions (at least) can be adapted to the properties of the inner product functions via the Kernel trick. In general, an utility function can be classified within the broad range of von Neumann Morgenstern (sequence of) utility functions [34] with pointwise convergence or with almost anywhere convergence. This paper utilizes the inner products as an alternative to utility functions in Economic Theory, including the uniqueness of optimal choices supported by budget constraints that are also inner products in the price and commodity space. Suppose, the budget set facing a consumer is given by: $Bp_w = \{x | px \leq w\}$ where, $w =$ income or wealth; the budget frontier (line or hyperplane) is given by $\overline{B} = \{x | px = w\}$. If p is orthogonal to a budget frontier $\overline{B}$, then if, $x, y \in \overline{B}$, consequently the inner product is $p(y - x)$, where, $p(y - x) = py - px = w - w = 0$. In terms of a classification problem, the linear (and non-linear, where the inner product is raised to a power of integers) frontier (or hyperplane) is quite useful. In all unconstrained cases, the utility function itself may operate as the classifier. Indeed, as mentioned above, for von Neumann-Morgenstern utility functions, the degree of risk aversion (in the case of Hyperbolic Absolute Risk Aversion, HARA, is the risk tolerance instead) can classify elements according to distinct risk-taking abilities, ex-post. In related literature, masses of individuals who cannot be distinguished ex-ante based on the risk-taking abilities or that such as assumption would be an imposition, utility functions with comparable arguments can lead to ex-post categorization based on one or more critical parameters. These arguments from Economics lay the foundation for our SVM Utility Kernel [35].

The Utility Kernel: The Utility kernel for any two inputs x_i, z_i is stated as: $K(x_i, z_i) = k_0 + k_1 < x_i, z_i >^{\alpha}$. Here, α k_0, k_1 are kernel parameters. A kernel must satisfy Mercer's theorem as follows.

Theorem 1: Mercer's Theorem: *If K is a continuous symmetric function such that the integral operator $L_K : L^2 \rightarrow L^2$, defined as $L_Kg(x) = \int K(x, y)g(y)dy$ is positive, then it means that for all functions $g \in L^2$, the condition $\int\int K(x, y)g(x)g(y)dxdy \geq 0$ stands true.*

Sketch of Proof: Utility kernel $K(x, z)$ defined as $K(x_i, z_i) = k_0 + k_1 < x_i, z_i >^{\alpha}$ is a kernel if $K(x, z)$ is symmetric, continuous, and positive semi definite. Since the inner product of inputs x_i, z_i $(\forall x_1, ..., x_n \in \Re$ and $\forall z_1, ...$ $z_n \in \Re)$ is symmetric, the $K(x_i, z_i)$ is symmetric (this implies K(x,z)=K(z,x)) and also continuous.

Let $K(x, y)$ be a continuous, non-negative definite, symmetric function of variables x and y. Let f be a function such that all $f \in L^2$, where L^2 is square-integrable function. Moreover, the integral operator L_K (when $L_K : L^2 \rightarrow L^2$) with kernel K is defined as

$$L_Kf(x) = \int K(x, y)f(y)dy \tag{1}$$

We say that K satisfies Mercer's conditions, for all values of f belonging to L^2, *iff*

$$\int\int K(x, y)f(x)f(y)dxdy \geq 0 \tag{2}$$

Assuming that it is an n dimensional space, we re-write the proposed Utility kernel -

$$K(x, z) = k_0 + k_1(x^Tz)^{\alpha} = k_0 + k_1\left(\sum_{k=1}^{n} x_kz_k\right)^{\alpha} \tag{3}$$

Applying binomial series expansion, given as -

$$(x+z)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} z^k \tag{4}$$

on equation (3). We introduce integers $p_1, \ldots, p_{n-1}$, such that $p_i > p_{i+1}$ for all i ranging from 1 to $n-2$.

$$= k_0 + k_1 \sum_{p_1}^{\alpha} \binom{\alpha}{p_1} (x_1 z_1)^{\alpha - p_1} \times \sum_{p_2}^{p_1} \binom{p_1}{p_2} (x_2 z_2)^{p_1 - p_1} .... \times \sum_{p_{n-1}}^{p_{n-2}} \binom{p_{n-2}}{p_{n-1}} (x_{n-1} z_{n-1})^{p_1 - p_1} \times (x_n z_n)^{p_{n-1}} \tag{5}$$

Rearranging the order of the sums and integral, and exploiting the property of the square integral operator, L^2, we solve the above equation. The property indicates that, if f is a square integral function defined as $f : \Re \to C$ then, $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$, $\forall f \in L^2(\chi)$ where $\chi \subseteq \Re^n$. We also know that the space L^2 which is square integral function, is an example of Hilbert space, with inner product defined as

$$< f, f > = \int f(x) f(x) dx = \int f^2(x) dx = \int |f(x))|^2 dx \tag{6}$$

We obtain -

$$= k_0 + k_1 \sum_{p_1}^{\alpha} \sum_{p_2}^{p_1} \ldots \sum_{p_{n-1}}^{p_{n-2}} \frac{\alpha!}{(c_1)!(c_2)!....(c_{n-1})!} \left(\int (x_1)^{c_1} \ (x_2)^{c_2} ... (x_n)^{c_n} f(x_1, x_2, ..) dx_1, dx_2, .. \right)^2$$

Assuming $k_0 > 0$ and $k_1 > 0$, we conclude that the above equation has square term and is non-negative, thereby providing enough evidence of Utility kernel being positive semi-definite. Hence it meets Mercer's all three conditions. Note, we assume, additionally, $(\alpha - p_1) = c_1$, $(p_1 - p_2) = c_2$, $(p_{j-2} - p_{j-1}) = c_{j-1}$ and $p_j = c_j$, for $j = 2, \ldots, n$. Please see Section 1, supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2022.3198879 for additional details.

Remark: It can be shown easily that by keeping values of k_0 and k_1 as 1, the polynomial kernel manifests as a special case of utility kernel with extra dimensions in transformed space (See Section 3.3, supplementary file, available online) that shows up due to the structure of the kernels. We test the Utility Kernel empirically on data and report the results in Section 7.

Theorem 2: The upper bound on the generalization error of an optimal hyperplane is a smooth function of utility kernel parameters (k_0, k_1, α).

Proof: Please see Section 2 in the supplementary file, available online.

Implication of Theorem 2: To show that generalization error is a continuous function of kernel parameters, we start by investigating the optimal margin 'm' and its continuity with respect to (k_0, k_1, α). We use the implicit function theorem to prove that the margin 'm' is a continuous function of kernel parameters and assert that the margin is sensitive to k_0, k_1, α and thus establish the significance of the hyper-parameters in the utility kernel to obtain better generalization errors. We have shown through various plots (Figs. 1 and 2, supplementary file, available online Section2) that generalization error varies smoothly by varying the Utility kernel parameters.

Theorem 3: VC Dimension of Utility Kernel: Let $F = \{f | f(x) = k_1 < x, x >^{\alpha} + k_0, w \in \mathbb{R}^d, k_0, k_1 \in \mathbb{R}\}$. Then, H is the set of all linear classifiers,

$$H = \{1_{k_1 < w, x >^{\alpha} + k_0 \geq 0} | w \in \mathbb{R}^d, k_0, k_1 \in \mathbb{R}\}$$

$$and \ dim(F) = d + 1 \tag{7}$$

Proof: Please see Section 3.1 in the supplementary file, available online.

Implication of Theorem 3: This result establishes a bound for classifiers in our case, Utility kernel which produces higher dimensional linear embedding from non-linearly separable data. Such a bound exists, and thus a notion of the capacity of such binary (linear) classifiers in higher dimensions is obtained.

Theorem 4: Upper bound of VC dimensions

Assume training examples $S = \{x_1, x_2 ... x_l\} \in \Re^d$ when transformed by function ϕ, the space of the transformed features is bounded by smallest hyper-sphere with radius R and center C. The VC dimension VC_K for kernel K is bounded by the following inequality -

$$VC_K \leq min\left(\left[\frac{R^2 w^2}{4} \right], d \right) + 1$$

Proof: Please see Section 3.2 in the supplementary file, available online.

Implication of theorem 4: We obtain an upper bound for the capacity of the classifier. We achieve that by minimizing $4\frac{R^2}{m^2}$, or essentially $R^2 \|w\|^2$. The term $R^2 \|w\|^2$ reflects the capacity of the kernel map. We observe that, for smaller R^2, the VC capacity is low, h is low, and eventually, a better generalization performance of the kernel is assured.

Theorem 5: Utility Kernel is an Universal Approximator.

Proof: Please refer Section 4 in the Supplementary file, available online.

Implication of Theorem 5: The theorem shows that the Utility kernel can approximate any non-linear continuous function well. Additionally, we show that the Utility Kernel is a universal approximator and is a universal kernel (refer to Section 4.1 supplementary file, available online).

5 Computing Parameters of Utility Kernel Using Exponentially Weighted Momentum Based Particle Swarm Optimization (EMPSO)

PSO [36] is a popular, highly robust, conceptually simple, stochastic optimization technique that simulates the social behaviour seen in bird flocks or swarm. In this approach, particles search for optimal solution in d-dimensional search space. Each particle is represented as x_{ij} and its velocity is indicated as v_{ij}, where i is the index of i^{th} particle in j^{th} dimension. The best position of the i^{th} particle is x_i^{pbest}

and the best position of the swarm is x^{gbest}. Another variant of PSO, Exponentially Weighted Momentum PSO (EMPSO) [37] involves momentum term, M_i for i^{th} particle. With every iteration, t, the formula to update velocity, position and momentum of a i^{th} particle for the next iteration, $t+1$ are given as,

$$v_{ij}(t+1) = M_{ij}(t) + c_1 r_1 (x_{ij}^{pbest}(t) - x_{ij}(t)) + c_2 r_2 (x_{ij}^{gbest}(t) - x_{ij}(t)) \quad (8)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (9)$$

$$M_{ij}(t+1) = \beta M_{ij}(t) + (1-\beta) v_{ij}(t) \quad (10)$$

where, β is the momentum factor, M_{ij} is the momentum for the i^{th} particle in j^{th} dimension, c_1, c_2 are the learning factors (non-negative), and r_1, r_2 are the random values distributed uniformly between 0 and 1. The EMPSO restricts the particles to search for solutions within a range of [0, 10] and prevents them from moving away from the search space. The fitness function used in the algorithm is the loss obtained from every particle after being trained by kernel SVM. So, every particle will generate loss (or error) based on its value being used in the SVM kernel, and the particle (assume x_b) for which the predicted error is minimum is assumed to be at the best position. Subsequently, the positions and velocities of other particles are updated to move towards x_b. The process is repeated till the termination criterion of achieving minimum fitness function (error) or maximum number of iterations is reached.

EMPSO algorithm finds the optimal kernel parameters (k_0, k_1 and α) by initializing a population of K random particles in 3-dimensional space. Their positions, velocities and momentum are computed as per Equations (8), (9) and (10). Function *SVM-kernel*(x_{ij}) is the execution of utility kernel on the input data, D. The parameters of utility kernel are utilized from x_{ij} (i.e., $x_{i0} = k_0, x_{i1} = k_1, x_{i2} = \alpha$). For every particle, $k \in K$, the fitness function at t^{th} iteration given by sum of the cross entropy loss over the dataset D-

$$f_k^t = -\sum_D \sum_d y_k^t log(\hat{y}_k^t)$$

where d is the number of classes, y_k^t is the target output, and $\hat{y}_k^t$ is the predicted output. During the course of execution, max_iter iterations are executed and, for t^{th} iteration ($t \in max\_iter$), the best particle position is the one for which f_k^t is minimum. While updating position, we use round( ) function with x_{i2} to get the nearest integer value since α (the degree in the utility kernel formulation) cannot be a real value. EMPSO yields the optimal hyperparameter values which are used in the experimentation. These values of k_0, k_1, α are $0.1, 1000, 4$ for CLN modality and 0.1, 10, 5 for all other modalities and their combinations.

6 Experimentation

6.1 Dataset

The dataset used in this study is taken from "The Cancer Genome Atlas Program for Breast Cancer (TCGA-BRCA)[2]. The dataset consisted of information from six modalities (i.e., Clinical details of 1097 patients, mRNASeq of 1093 patients, Copy number variations of 1089 patients, Gene-Methylation of 1097 patients, miRSeq of 1078 patients, and Histopathological Whole Slide Images of 1089 patients). The raw dataset is incomplete-multiview data with certain patient information unavailable in various modalities. For our study, we have selected the complete multi-view data among all the modalities using intersection, and the final data consists of 1068 patients from clinical (CLN), mRNA-Seq (EXP), copy number variation (CNV), and histopathological whole slide images (WSI) modalities. We further studied the distribution of patients in terms of survival years as shown in Fig. 3. With the help of survival distribution, our problem is defined as the binary classification task of classifying patients as long-term survivors (marked as 0) and short-term survivors (marked as 1) based on various cut-off years in the range of five to nine years.

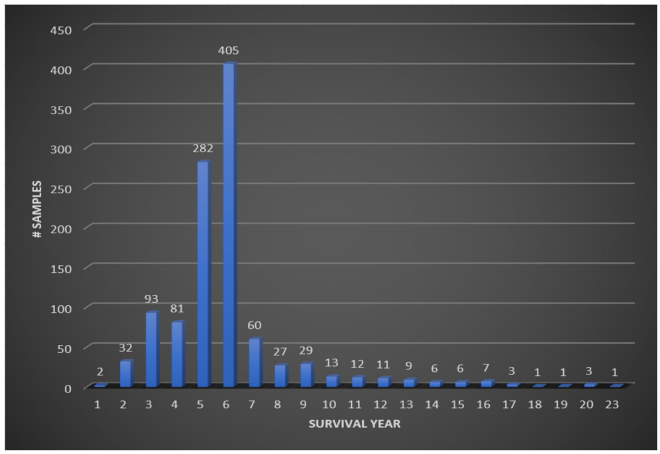

Fig. 3. TCGA-BRCA patient distribution in terms of survival years.

6.1.1 Genetic and Clinical Data Pre-Processing

The missing values of EXP profile data and CNV profile data are estimated using the weighted nearest neighbour approach [38]. In this technique, for estimating the missing value of gene 'X', it finds 'K' other genes with similar gene expressions to find the weighted average of 'K' similar genes as an estimate for the missing value of gene 'X'. The weighted average is calculated using the euclidean distance similarity measures between gene 'X' and other 'K' similar genes. Following the study by *Gevaert* et al. [12] over microarray data, we have used the already background corrected, normalized and log-transformed 20000 expression values (approximately) per patient and subsequently discretized them according to two thresholds. The gene with high variance goes to high threshold, and low variance belongs to a low threshold with values -1, 1, and 0 as under-expressed genes, over-expressed genes, and baseline-genes, respectively. With five discrete values (-2, -1, 0, 1, 2), the CNV features are employed as available. EXP and CNV profiles have large feature spaces containing genetic details of 20532 and 19729 genes, respectively. The dataset having feature space larger than the number of samples may suffer from the curse of dimensionality [39] during machine learning training. Hence we decided to reduce the feature dimensions without losing

2. https://portal.gdc.cancer.gov/

the relevant information. For this, we borrowed the mRMR feature selection [40] [41] [42] technique used by *Arya et. al.* [23] over the same dataset, and selected 400 relevant genes from EXP modality followed by 200 important genes from CNV modality. The clinical profile used in the study includes ten initial follow-up level indicators ranging from the basic information of patients to some details related to tumors and cancer. The age attribute of clinical data is normalized in the range [0,1] using min-max normalization [10] and other categorical features are used as it is.

Algorithm 1. Parameters of EMPSO-SVM (P, K, n, t^{max}, $x_{ij}, v_{ij}, m_{ij}, x^{pbest,t}, x^{gbest,t}, f^t, c_1, c_2, r_1, r_2$, D)

Input: P-Particle swarm, K-size of the swarm, n-dimension of search space, x_{ij}, v_{ij}, m_{ij} -particle's position, velocity and momentum, $x^{pbest,t}$-best position (particle), $x^{gbest,t}$ -global best, c_1, c_2 -acceleration constant, D-input data
Result: optimized values of k_0, k_1, α
Initialize swarm positions $x_{ij}(i = 1...K, j = 1...n)$
for *each i in K* **do**
  for *each j in n* **do**
    Calculate v_{ij}, m_{ij} as initial velocity and momentum using Equations (8) and (10)
  end
end
set $t = 0$
for *each t in* t^{max} **do**
  `/* t signifies iterations */`
  Randomly choose r_{1t}, r_{2t}
  for *each particle k in K* **do**
    $prediction\_error\_array = prediction\_error\_array \bigcup prediction\_error$
    Evaluate fitness function f_k^t using SVM-kernel(x_k)
    $f_k^t \leftarrow$ SVM-kernel(x_k)
    $x^{gbest,t} \leftarrow f_{k,min}^t$
  end
  for *each particle k in K* **do**
    Find particle best position- x_k^{pbest}
    Update velocity v_k^{t+1}, and momentum m_k^{t+1} with respect to x_k^{pbest} and $x^{gbest,t}$ using equations (8) and (10)
    Compute x_k^{t+1} using equation (9)
  end
end

6.1.2 WSI Pre-Processing and Feature Extraction

To represent and encode WSIs, we need to develop machine learning methods that can effectively extract hidden informative features from WSIs. However, the high resolution of WSIs makes learning from them in their entirety difficult. Thus, there must be an element of stochastic sampling and filtering involved. In this work, we use a relatively simple approach to sample WSIs. We sample 256 $\times$ 256 pixel patches at the highest resolution using "PyHIST: A Histological Image Segmentation Tool" [43]. Then, we select the top 20% of the generated patches (or 40 patches) with highest RGB density as region of interests (ROIs); this ensures that 'non-representative' patches belonging to white-space are ignored, and densest tiles include more cells for further investigations [19]. These 40 ROIs represent, on average, 15% of the tissue region within the WSI. Each of these 40 patches is passed through a deeper but less complex CNN, ResNet152 [44] pre-trained over ImageNet dataset [45] to get the hidden informative features of 2048 dimensions from last hidden layer. Following the PCA as better dimensionality reduction technique for gene microarray data in the study of *Khademi et. al.* [15], we projected the extracted features into the feature space of 512 dimensions. Further, we concatenated embeddings of all 40 patches to generate the WSI features of size 20480 for each patient. In WSI case also, training of machine learning model can not happen due to the small number of samples and the large feature space. So, we further reduced the features to 600 dimensions with the help of PCA, while preserving 95% variance. Finally, we have the dataset of size 1068 $\times$ 600 for WSI modality.

The final dataset has 10, 400, 200, and 600 relevant features from CLN, EXP, CNV and WSI modality, respectively.

6.2 Experimental Setup

All the experimental analyses in this study are carried out using ubuntu 18.04.5 LTS system along with 11 GB of NVIDIA GeForce RTX 2080 Ti and 32 GB RAM. The coding setup is having keras-gpu 2.2.4 with tensorflow-gpu 1.14.0 in backend and python 3.6 bundled in anaconda environment The source code is available at [3].

The proposed novel SVM utility kernel is used for the binary classification task of breast cancer survival prediction using all the modalities. Our experimental analysis starts with uni-modal predictions and ends with multi-modal predictions. In this journey, we tested for all possible combinations of modalities such as bi-modal, tri-modal, and multi-modal. So, we have fifteen unique combinations of modalities (i.e., CLN, WSI, EXP, CNV as uni-modal; CLN-EXP, CLN-CNV, CLN-WSI, EXP-CNV, EXP-WSI, CNV-WSI as bi-modal; CLN-EXP-CNV, CLN-EXP-WSI, EXP-CNV-WSI, CNV-CLN-WSI as tri-modal; and CLN-WSI-EXP-CNV as multi-modal). Each of these combinations has experimented with various survival year cut-offs ranging from 5-years to 9-years. For every combination, we run 10-fold cross-validation and record the average of every metric for all the 10 folds. Additionally, we have repeated this exercise 20 times and reported the average (mean) of 20 runs for MCC, Cohen-kappa, Jaccard index, G-measure, and TPR/TNR. From Fig. 3, it is conspicuous that the class distribution is not balanced for any survival year cut-off. Hence, our training setup is designed by taking an equal number of samples from both classes and training the proposed SVM in 10 fold cross-validation framework. In the process of balancing classes using the sampling technique, few instances from the majority class are left out. To avoid the bias of the model towards chosen samples only, we propose the Bootstrap Minority Class Balancing (BMCB) method (see details of lemmas 1-3 and conjecture in the supplementary file, available online, Section 6) where we ensure that millions of subsets are created within the training set. Here, the majority class is down-sampled in each subset in such a way that the union of the subsets guarantees the inclusion of all instances from dataset without

3. GitHub: https://github.com/nikhilaryan92/utilitykernel

missing a single pattern of the data. We have tuned the regularization parameter (C) and hyper-parameters (k_0, k_1 and α) of the utility kernel as 500, 0.1, 1000, 4 for CLN modality and 1, 0.1, 10, 5 for all other modalities and their combinations, respectively. These values are obtained from the EMPSO algorithm discussed in Section 5. As a cross-validation exercise to check the optimal values of parameters received from EMPSO, we performed a method similar to a grid search. We froze values of k0 and k1 to small numbers (0.01 and 1.1) and iterated α from values 1 to 15 till it returned the best results. We got the best results for α=4 for the CLN modality and α=5 for other modalities. In the next step, k0 set to 0.01, and α was assigned its best value. The executions were performed by iterating k1 from 1 to 1000, and we found the best results at k1=1000 for CLN and, k1= 10 for the remaining modalities. Later on, values of k1 and α were frozen, and k0 was iterated. We got the best results at k0=0.1, thus leading us to the empirical validation for using the EMPSO as an optimization technique. It was also done for the polynomial kernel. The degree of the polynomial kernel was iterated from 2 to 6. However, in random forest-based classifiers, the number of decision trees can not be determined using a metaheuristic approach like the EMPSO. A small value of k_0 received from EMPSO ensures that similar samples of the same class are still different. Before applying the proposed kernel function in the role of classification, we validated that the features in this study are neither linearly separable nor correlated to each other.

To validate the efficacy and robustness of the utility kernel, we have experimented with other popular SVM kernel functions such as linear, polynomial, RBF, and sigmoid in the breast cancer survival prediction task. Along with SVM, we have also experimented with other machine learning methods like Naive Bayes, Decision tree, and Random Forest. Nowadays, deep neural networks are proposed to be superior to machine learning methods in the cancer survival prediction. So, we extended our experiments to use variational autoencoders as hidden feature extractors followed by deep neural networks as final classifiers. We also tried various transfer learning techniques using popular CNNs (i.e., VGG-Nets, ResNets, DenseNets, etc.). The complex architectures of deep neural networks and transfer learning-based CNNs were unable to learn breast cancer survival prediction from the given small-sized high dimensional, highly imbalanced dataset, which motivated us to explore the efficacy of various machine learning techniques, specifically SVMs.

7 Performance Evaluation and Results

In this section we have reported the experimental results obtained.

7.1 Performance Evaluation Measures

We performed our evaluation criteria ranging from conventional performance measures like accuracy, precision, sensitivity, specificity, and $F_{1-score}$ to Matthews Correlation Coefficient (MCC) $=\frac{tp\times tn-fp\times fn}{\sqrt{(tp+fp)\times(tp+fn)\times(tn+fp)\times(tn+fn)}}$, Cohen's Kappa (CP) $=\frac{2\times(tp\times tn-fn\times fp)}{(tp+fp)\times(fp+tn)+(tp+fn)\times(fn+tn)}$, G-measure (GM) $=\sqrt{precision\times recall}=\sqrt{\frac{tp}{tp+fp}\times\frac{tp}{tp+fn}}$, Jaccard index (JI) $=\frac{tp}{tp+fp+fn}$, TPR/TNR $=\frac{\frac{tp}{tp+fn}}{\frac{tn}{tn+fp}}$, and ROC-AUC, (the conventional, metrics such as $F_{1-score}$, accuracy, etc. are reported in the supplementary , available onlinefile, Section 9). Here, true positive (tp), true negative (tn), false positive (fp), and false negative (fn) denotes the four component of the confusion matrix.

The best models are selected after comparing MCC (Matthews Correlation Coefficient), Cohen's Kappa, G-measure, and Jaccard index for every classifier. If the classifier efficiently and correctly classifies the patients, then it scores higher in terms of all the performance indicators (MCC (Matthews Correlation Coefficient), Cohen's Kappa, G-measure, and Jaccard index). We have also used TPR/TNR as an additional performance measure which tells about the classifier's ability to balance between positive and negative class classification. If the ratio of TPR and TNR is too high (more than one), then the classifier is biased towards positive class identification. If it is too low (close to zero), then the classifier is only able to identify negative classes properly.

7.2 Results

In the results section, we have presented a detailed and comparative analysis of the SVM utility kernel concerning other kernels along with machine learning and deep learning architectures. The results have been compared on varying survival year cut-offs. The very first analysis is performed on a 5-year survival cut-off with 20% samples as short-term survivors and 80% as long-term survivors, followed by 6-year, 7-year, 8-year, and 9-year survival cut-off with the class distribution of 45%-55%, 83%-17%, 89%-11% and 92%-8%, respectively. Here, the 6-year survival cut-off is balanced to some extent, while others are highly imbalanced in class distribution. Figures 1a, 1b, 1c, 1d from supplementary file, available online and Table 1 in this manuscript illustrate the 5-year breast cancer survival prediction based comparative analyses of Utility kernel with other popular SVM kernels, Naive Bayes classifier, Decision tree and Random forest in terms of Matthew's correlation coefficient, cohen's kappa, G-measure and Jaccard-index, respectively. Utility kernel scores the highest MCC of 94.09% for uni-modal gene-expression profile, 94.31% for bi-modal gene-expression integrated with clinical profile, and 91.34% for tri-modal clinical profile integrated with gene-expression and copy number variation. The same uni-modal, bi-modal, and tri-modal profiles scored the highest cohen's kappa values of 93.96%, 94.19%, and 90.65%; high-

TABLE 1
Mean Evaluation Measures for Best Possible Breast Cancer Survival Prediction Models at 5-Year Survival

Modality	Best Classifiers	MCC	CP	$\frac{TPR}{TNR}$	GM	JI
CLN	Random Forest	0.6537	0.5983	1.2455	0.7311	0.5363
WSI	Polynomial SVM	0.7073	0.671	0.8758	0.7421	0.5864
EXP	**Utility kernel SVM**	**0.9409**	**0.9396**	**0.9363**	**0.9512**	**0.9065**
CNV	Random Forest	0.4183	0.2772	1.7935	0.5549	0.3186
CLN-EXP	**Utility kernel SVM**	**0.9431**	**0.9419**	**0.9327**	**0.9531**	**0.9097**
CLN-CNV	Random Forest	0.5981	0.5239	1.2915	0.6855	0.4769
CLN-WSI	Polynomial SVM	0.683	0.6388	0.9662	0.7264	0.5633
EXP-CNV	**Utility kernel SVM**	**0.9066**	**0.8989**	**0.9762**	**0.9249**	**0.8630**
EXP-WSI	Polynomial SVM	0.7052	0.6677	0.8869	0.7408	0.5843
CNV-WSI	Polynomial SVM	0.6989	0.6596	0.9105	0.7366	0.577
CLN-EXP-CNV	**Utility kernel SVM**	**0.9134**	**0.9065**	**0.9819**	**0.9309**	**0.8745**
CLN-EXP-WSI	Polynomial SVM	0.7118	0.6774	0.8691	0.7455	0.5917
EXP-CNV-WSI	Polynomial SVM	0.6946	0.653	0.9356	0.7351	0.5729
CNV-CLN-WSI	Polynomial SVM	0.709	0.6727	0.9075	0.7467	0.5896
CLN-WSI-EXP-CNV	Polynomial SVM	0.7073	0.6708	0.8915	0.7433	0.5869

Mean performance metrics (of 20 runs) are reported.

TABLE 2
Mean Evaluation Measures for Best Possible Breast Cancer Survival Prediction Models at 6-Year Survival

Modality	Best Classifiers	MCC	CP	$\frac{TPR}{TNR}$	GM	JI
CLN	Random Forest	0.8611	0.8503	1.0663	0.9215	0.8532
WSI	**Utility kernel SVM**	**0.8913**	**0.8845**	**0.9888**	**0.9372**	**0.8816**
EXP	**Utility kernel SVM**	**0.9185**	**0.9145**	**1.0186**	**0.9547**	**0.9129**
CNV	Random Forest	0.6745	0.614	1.4124	0.8237	0.6874
CLN-EXP	**Utility kernel SVM**	**0.9127**	**0.9082**	**1.0022**	**0.9514**	**0.9062**
CLN-CNV	Random Forest	0.8847	0.8774	1.0542	0.9353	0.8776
CLN-WSI	**Utility kernel SVM**	**0.8919**	**0.8853**	**0.9864**	**0.9375**	**0.8823**
EXP-CNV	Random Forest	0.8974	0.8917	1.0697	0.9435	0.8917
EXP-WSI	**Utility kernel SVM**	**0.8995**	**0.8938**	**0.9794**	**0.9418**	**0.8899**
CNV-WSI	**Utility kernel SVM**	**0.8828**	**0.8749**	**1.0112**	**0.9331**	**0.874**
CLN-EXP-CNV	**Utility kernel SVM**	**0.9076**	**0.9031**	**1.0732**	**0.9496**	**0.9026**
CLN-EXP-WSI	**Utility kernel SVM**	**0.9003**	**0.8948**	**0.9731**	**0.9421**	**0.8905**
EXP-CNV-WSI	**Utility kernel SVM**	**0.8914**	**0.8847**	**0.9872**	**0.9372**	**0.8817**
CNV-CLN-WSI	**Utility kernel SVM**	**0.8968**	**0.8909**	**0.9779**	**0.9401**	**0.8869**
CLN-WSI-EXP-CNV	**Utility kernel SVM**	**0.9002**	**0.8947**	**0.9734**	**0.9421**	**0.8904**

Mean performance metrics (of 20 runs) are reported.

TABLE 4
Mean Evaluation Measures for Best Possible Breast Cancer Survival Prediction Models at 8-Year Survival

Modality	Best Classifiers	MCC	CP	$\frac{TPR}{TNR}$	GM	JI
CLN	Random Forest	0.9099	0.9069	1.1225	0.9844	0.9692
WSI	**Utility kernel SVM**	**0.9297**	**0.9281**	**1.0638**	**0.9875**	**0.9754**
EXP	**Utility kernel SVM**	**0.9437**	**0.9423**	**1.0932**	**0.9903**	**0.9808**
CNV	Random Forest	0.7905	0.7704	1.4713	0.9663	0.934
CLN-EXP	**Utility kernel SVM**	**0.943**	**0.9415**	**1.0971**	**0.9902**	**0.9806**
CLN-CNV	Random Forest	0.9375	0.9357	1.1087	0.9893	0.9788
CLN-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
EXP-CNV	Random Forest	0.9404	0.9387	1.1078	0.9898	0.9798
EXP-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
CNV-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
CLN-EXP-CNV	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CLN-EXP-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
EXP-CNV-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CNV-CLN-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CLN-WSI-EXP-CNV	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793

Mean performance metrics (of 20 runs) are reported.

est G-measure values of 95.12%, 95.31%, and 93.09%; highest Jaccard-index values of 90.65%, 90.97%, and 87.45%, respectively. If we consider 6-year breast cancer survival prediction based comparative analysis, then Figures 1e, 1f, 1g, 1h in the supplementary file, available online and Table 2 in this manuscript illustrate the effectiveness of Utility kernel. From Table 2, we can observe that the Utility kernel is being selected eleven times as the best performing model out of fifteen possible combinations of modalities. The highest MCC, cohen's kappa, G-measure, and Jaccard-index obtained by our Utility kernel for the uni-modal gene-expression profile are 91.85%, 91.45%, 95.47%, and 91.29%; for bi-modal clinical profile integrated with gene-expression are 91.27%, 90.82%, 95.14%, and 90.62%; for tri-modal clinical profile integrated with gene-expression and copy number variation are 90.76%, 90.31%, 94.96%, and 90.26%, respectively. Figures 1i, 1j, 1k, 1l in the supplementary file, available online and Table 3 in the manuscript demonstrate the 7-year breast cancer survival prediction based comparative analysis of Utility kernel with respect to other popular SVM kernels and some popular machine learning algorithms. Here also Utility kernel achieved the highest scores for all the evaluation criteria mentioned earlier. The highest MCC, Cohen's kappa, G-measure, and Jaccard-index among all possible combinations of modalities are 94.37%, 94.23%, 99.03%, and 98.08%, respectively, and these are scored by Utility kernel. If we observe the breast cancer survival prediction at 8-year survival cut-offs with very high class imbalance from Figures 1m, 1n, 1o, 1p in the supplementary file, available online and Table 4 in the manuscript, where 89% of the samples are short-term survivors then Utility kernel is giving comparable results with random forest classifier and RBF SVM in breast cancer survival prediction. But, analysis of Figures 1q, 1r, 1s, 1t in the supplementary file, available online and Table 5 in the manuscript for 9-year survival prediction shows the failure of Utility kernel SVM over RF and RBF SVM. One possible reason could be RF is an ensemble and, in the absence of label noise and attribute noise, is one of the best classifiers. To find the root cause for the failure of the utility kernel SVM, we performed a thorough analysis of the data-set for the 9-year survival task, and the following observations have been noticed:

1) The majority of the samples belonging to different classes (i.e., class 0 and class 1) of 9-year survival are highly similar to each other. Their cosine similarities are higher than 0.999, but they belong to the same class for 5 and 6-year survival.
2) The majority of the samples belonging to same class of 9-year survival are very much dissimilar. Their cosine similarities are lesser than 0.20, but they belong to different classes for 5 and 6-year survival.

As we already know from the working principle of SVM, it tries to find the best possible hyperplane, which acts as a decision boundary between the points belonging to different classes. If the points belonging to a separate class are highly cosine similar (approximately equal to 1) then the cosine distance between these points becomes zero, which makes the points identical, and confuses the SVM in the process of finding the decision boundary for these points. Similarly, different points of the same class are low cosine

TABLE 3
Mean Evaluation Measures for Best Possible Breast Cancer Survival Prediction Models at 7-Year Survival

Modality	Best Classifiers	MCC	CP	$\frac{TPR}{TNR}$	GM	JI
CLN	Random Forest	0.9099	0.9069	1.1225	0.9844	0.9692
WSI	**Utility kernel SVM**	**0.9297**	**0.9281**	**1.0638**	**0.9875**	**0.9754**
EXP	**Utility kernel SVM**	**0.9437**	**0.9423**	**1.0932**	**0.9903**	**0.9808**
CNV	Random Forest	0.7905	0.7704	1.4713	0.9663	0.934
CLN-EXP	**Utility kernel SVM**	**0.943**	**0.9415**	**1.0971**	**0.9902**	**0.9806**
CLN-CNV	Random Forest	0.9375	0.9357	1.1087	0.9893	0.9788
CLN-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
EXP-CNV	Random Forest	0.9404	0.9387	1.1078	0.9898	0.9798
EXP-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
CNV-WSI	Random Forest	0.9378	0.9358	1.1131	0.9894	0.9789
CLN-EXP-CNV	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CLN-EXP-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
EXP-CNV-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CNV-CLN-WSI	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793
CLN-WSI-EXP-CNV	Random Forest	0.9391	0.9372	1.1105	0.9896	0.9793

Mean performance metrics (of 20 runs) are reported.

TABLE 5
Mean Evaluation Measures for Best Classifiers Breast Cancer Survival Prediction Models at 9-Year Survival

Modality	Best Classifiers	MCC	CP	$\frac{TPR}{TNR}$	GM	JI
CLN	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
EXP	RBF SVM	0.9507	0.9495	1.0962	0.9954	0.9908
CNV	Random Forest	0.8005	0.7829	1.4834	0.9826	0.9656
CLN-EXP	RBF SVM	0.9505	0.9492	1.0961	0.9954	0.9908
CLN-CNV	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
CLN-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
EXP-CNV	RBF SVM	0.948	0.9467	1.0989	0.9951	0.9903
EXP-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
CNV-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
CLN-EXP-CNV	RBF SVM	0.9479	0.9466	1.1009	0.9951	0.9903
CLN-EXP-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
EXP-CNV-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
CNV-CLN-WSI	Random Forest	0.9312	0.9295	1.1152	0.9936	0.9872
CLN-WSI-EXP-CNV	Random Forest	0.9392	0.9374	1.1209	0.9944	0.9887

Mean performance metrics (of 20 runs) are reported

similar (tending toward zero), the cosine distance becomes large, drifting these points to another side of the decision boundary. Similar cases are also observed for the 7 and 8-year survival. Hence, the proposed utility kernel is not able to showcase its efficacy for these particular scenarios. RBF SVM is usually reliable because of its ability in finding similarities/dissimilarities in class distribution and is therefore used as a popular SVM kernel. It is a shift-invariant kernel, i.e., it is invariant to translation and is now used in Support Vector Neural Networks [46]. Shift-invariance implies the kernel yields the same value $K(x, y)$ for $K(x + c, y + c)$, where c may be vector-valued of dimension to match the inputs. The linear kernel/utility kernel does not have the stationary property. The single-parameter version of the RBF kernel has the property that it is isotropic, i.e., the scaling by Y occurs by the same amount in all directions. It can be easily generalized by slight manipulation in the functional form to $K(x, y) = e^{-(x-y)^T} D(x - y)$ where D is a positive semi-definite matrix.

Generalization: We have compared the results of the Utility kernel with the other kernels over 13 benchmark data sets (Supplementary file, available online, Section 7). The values of MCC, Cohen Kappa, G-measure and Jacard-index are reported in the Table (supplementary file, available online, Table 4). The parameters of polynomial kernel (α) are learned from EMPSO, and the results illustrate that the performance of the utility kernel is either better (on 9 out of 13 data sets) or at-par (on 4 datasets) with the other kernels. Additionally, we observe that the utility kernel outperforms the current SoTA method (including Adam, Adadelta, Adagrad etc.) and AdaSwarm [47] in the following data sets: Wisconsin Breast Cancer, Wheat Seeds, Heart Disease, Wine, Car Evaluation) and performs nearly similar on other data sets.

Remark: *Additional experimental results* using stratified cross-validation and sample weighting scheme, where classes are not balanced, are also reported in Supplementary file, available online (Section 5 and 6). These results empirically establish the superiority of the proposed kernel.

7.3 Statistical Significance Test

The prediction measures obtained by the utility kernel are statistically significant, i.e., these results have not occurred by chance. To validate the statistical significance, we have performed two sample two-sided (independent sample two-sided) t-test over the area under the ROC of utility kernel SVM and already existing SVM classifiers at various survival year cut-offs. The p-values for all possible modalities (i.e., uni-modal, bi-modal, tri-modal, and multi-modal) at 5-year survival are less than 0.05 (Please refer to table 11, supplementary file, available online). Hence, we can discard the null hypothesis (mean of both groups are identical), and conclude that the performance of our algorithm is statistically significant. As per Bonferroni analysis, Utility kernel outperforms the other rivalry machine learning classifiers (RF and Polynomial SVM) twenty and eighteen times in twenty independent runs of train test setup for 5-year and 6-year survivals, respectively. So, we can rely on the efficacy of the proposed utility kernel without any reasonable doubt.

8 Conclusion

We conclude this study with an interpretable and flexible utility kernel-based SVM for breast cancer survival prediction. The effectiveness of multi-modal data and inference in the correct prediction of breast cancer is established empirically. Section 8 in the supplementary file, available online elaborates on this and a case study is presented. Utility kernel has three hyper-parameters, k_0, k_1, and α, which make it flexible enough to fit on the integration of complex and heterogeneous data from various input sources like genomics, pathology, and tissue images. Along with the theoretical framework, we have also proved the empirical efficacy of the proposed utility kernel using TCGA-BRCA dataset by gaining the highest performance measures *for unconventional (Section 7, main text) and conventional (Section 9, supplementary file, available online)* metrics among other popular machine learning and deep learning architectures. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation [48]. If we use the sample weighting technique [Inverse of Number of Samples (INS), Inverse of Square Root of Number of Samples (ISNS)], the utility kernel is the best performing kernel across all uni-modal (except CLN) and multi-modal combinations for all survival year cut-offs. Please see Tables 2 and 3 of supplementary file, available online for the detailed analysis of MCC, Cohen's Kappa, G-measure, and Jaccard-index for the utility kernel as the best classifier. In case of the training technique (stratified 10-fold cross-validation without bootstrap minority class balancing) the utility kernel is the best classifier for ten out of fifteen combinations of modalities for a 6-year survival cut-off (only case where the classes are reasonably balanced). It outperforms all other kernels and random forest for multi-modality (CLN-WSI-EXP-CNV) combination (Please see Table 1, supplementary file, available online).

In the case of bootstrap minority class balancing (BMCB) for high-class imbalance, the 5-year cutoff-based utility kernel has outperformed all the other kernels and random forest classifier with the highest MCC value for EXP, CLN-EXP, EXP-CNV, and CLN-EXP-CNV among all possible combination of modalities. However, if we consider other, more conventional performance metrics, Utility Kernel remains the State-of-the-Art. In a few unis and multi-modality combinations, random forest and polynomial kernel seem to perform better in conventional metrics, but the highest MCC of the polynomial kernel and random is still worse than the MCC of the utility kernel. The 6-year cut-off-based utility kernel has been the best classifier with the highest MCC for eleven (WSI, EXP, CLN-EXP, CLN-WSI, EXP-WSI, CNV-WSI, CLN-EXP-CNV, CLN-EXP-WSI, EXP-CNV-WSI, CNV-CLN-WSI, and multi-modal) combination of modalities. Similarly, for a 7-year cut-off, utility kernel SVM is best performing with the highest MCC value of 0.9437 and it leads the second-best classifier (random forest) by 0.33%. If we refer to Table 4 of the main text, then the utility kernel SVM and random forest have comparable performance for 8-year survival cut-offs with equal MCC, cohen's kappa, G-measure, and Jaccard-index. The 9-year survival classification is not supporting the utility kernel as

the best classifier. Please see Table 1, 3, 2 of the main text for the highest obtained unconventional performance measures, which correspond to the best MCC value of the utility kernel. This utility kernel-based SVM is computationally efficient in comparison to advanced deep learning architectures. It does not require powerful resources for training and prediction. In the future, the proposed utility kernel can be applied to the survival estimation of other deadly diseases. Hence, it will be a savior for patients and oncologists for the clinical management of cancer and other life-threatening diseases and their prognosis.

Since the Kernel has more than one parameter to tune and the generalization error is a continuous function of the kernel parameters, it may be cumbersome to obtain the 'ideal' parameters for optimal performance. As such, a theoretical relationship guiding us to suitable kernel parameter values from the error bound would be useful to explore. Since the kernel trick works for a set of possible embeddings, its difficult to know apriori if the Kernel will work efficiently across diverse data sets. Additionally, the C parameter, which is very important for controlling the overfitting, needs to be optimized. In future we will also try to incorporate other modalities like mammography, gene-methylation and miRSeq for extracting features and then a multi-modal fusion model will be developed for breast cancer prognosis prediction.

ACKNOWLEDGMENT

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

REFERENCES

[1] G. M. Clark, "Do we really need prognostic factors for breast cancer?," *Breast Cancer Res. Treat.*, vol. 30, no. 2, pp. 117–126, 1994.

[2] L. R. Martin, S. L. Williams, K. B. Haskard, and M. R. Dimatteo, "The challenge of patient adherence," *Therapeutics Clin. Risk Manage.*, vol. 1, no. 3, pp. 189–199, Sep. 2005.

[3] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, Apr. 2012.

[4] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (TCGA): An immeasurable source of knowledge," *Współczesna Onkologia*, vol. 1A, pp. 68–77, 2015. [Online]. Available: http://www.termedia.pl/doi/10.5114/wo.2014.47136

[5] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0933365704001010

[6] K. Polyak, "Heterogeneity in breast cancer," *J. Clin. Investigation*, vol. 121, no. 10, pp. 3786–3788, Oct. 2011. [Online]. Available: http://www.jci.org/articles/view/60534

[7] Z. Obermeyer and E. J. Emanuel, "Predicting the future – Big Data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMp1606181

[8] L. J. van 't Veer et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002. [Online]. Available: http://www.nature.com/articles/415530a

[9] D. M. Abd El-Rehim et al., "High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses," *Int. J. Cancer*, vol. 116, no. 3, pp. 340–350, Sep. 2005.

[10] M. J. van de Vijver et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England J. Med.*, vol. 347, no. 25, pp. 1999–2009, Dec. 2002.

[11] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," in *Proc. 5th Int. Conf. BioMed. Eng. Inform.*, 2012, pp. 928–931.

[12] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–190, Jul. 2006.

[13] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 06, no. 05, pp. 551–560, 2013. [Online]. Available: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/jbise.2013.65070

[14] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, no. 1, pp. 30–37, Jan. 2007.

[15] M. Khademi and N. S. Nedialkov, "Probabilistic graphical models and deep belief networks for prognosis of breast cancer," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl.*, 2015, pp. 727–732.

[16] W. K. Moon et al., "Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images," *Comput. Methods Prog. Biomed.*, vol. 146, pp. 143–150, Jul. 2017.

[17] H. Wang, F. Xing, H. Su, A. Stromberg, and L. Yang, "Novel image markers for non-small cell lung cancer classification and survival prediction," *BMC Bioinf.*, vol. 15, no. 1, Dec. 2014, Art. no. 310. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471–2105-15-310

[18] X. Zhu et al., "Lung cancer survival prediction from pathological images and genetic data – An integration study," in *Proc. IEEE 13th Int. Symp. Biomed. Imag.*, 2016, pp. 1173–1176.

[19] K.-H. Yu et al., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nat. Commun.*, vol. 7, no. 1, Nov. 2016, Art. no. 12474. [Online]. Available: http://www.nature.com/articles/ncomms12474

[20] P. Dutta, A. P. Patra, and S. Saha, "DeePROG: Deep attention-based model for diseased gene prognosis by fusing multi-omics data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jun. 24, 2021, doi: 10.1109/TCBB.2021.3090302.

[21] S. J. Giri, P. Dutta, P. Halani, and S. Saha, "MultiPredGO: Deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1832–1838, May 2021.

[22] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, May/Jun. 2019.

[23] N. Arya and S. Saha, "Multi-modal classification for human breast cancer prognosis prediction: Proposal of deep-learning based stacked ensemble model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 1032–1041, Mar./Apr. 2022.

[24] N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," *Knowl.-Based Syst.*, vol. 221, 2021, Art. no. 106965. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121002288

[25] D. Sun, A. Li, B. Tang, and M. Wang, "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome," *Comput. Methods Prog. Biomed.*, vol. 161, pp. 45–53, Jul. 2018.

[26] B. Tang, A. Li, B. Li, and M. Wang, "CapSurv: Capsule network for survival analysis with whole slide pathological images," *IEEE Access*, vol. 7, pp. 26 022–26 030, 2019.

[27] N. Arya and S. Saha, "Generative incomplete multi-view prognosis predictor for breast cancer: GIMPP," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 2252–2263, Jul./Aug. 2022.

[28] Y.-F. Zhu, L.-F. Tian, Z.-Y. Mao, and L. Wei, "Mixtures of kernels for SVM modeling," in *Proc. Int. Conf. Natural Comput.*, 2005, pp. 601–607.

[29] W. Hai, "A support vector machine with a hybrid kernel and its application in underwater target recognition," *Tech. Acoust.*, vol. 154, pp. 127–133, 2005.
[30] G. Ai-ling, "Application of combined kernel SVM on network security risk evaluation," *Comput. Eng. Appl.*, vol. 45, pp. 122–125, 2009.
[31] M. Gönen and E. Alpaydın, "Localized algorithms for multiple kernel learning," *Pattern Recognit.*, vol. 46, no. 3, pp. 795–807, 2013.
[32] S. Cui-juan, "Support vector machine based k-type kernel function," *J. Huaihai Inst. Technol.*, vol. 4, pp. 4–7, 2006.
[33] V. Apostolidis-Afentoulis and K.-I. Lioufi, "SVM classification with linear and RBF kernels," 2015. [Online]. Available: http://www.academia.edu/13811676/SVM
[34] G. Chichilnisky, "Von neumann: Morgenstern utilities and cardinal preferences," *Math. Operations Res.*, vol. 10, no. 4, pp. 633–641, 1985. [Online]. Available: http://www.jstor.org/stable/3689431
[35] H. Beladi and S. Kar, "Skilled and unskilled immigrants and entrepreneurship in a developed country," *Rev. Develop. Econ.*, vol. 19, no. 3, pp. 666–682, 2015.
[36] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw.*, 1995, pp. 1942–1948.
[37] R. Mohapatra, S. Saha, C. A. C. Coello, A. Bhattacharya, S. S. Dhavala, and S. Saha, "Adaswarm: Augmenting gradient-based optimizers in deep learning with swarm intelligence," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 329–340, Apr. 2022.
[38] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Jun. 2001.
[39] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data," *Mol. Pharmaceutics*, vol. 13, no. 7, pp. 2524–2530, Jul. 2016. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.6b00248
[40] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[41] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.
[42] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, Apr. 2012.
[43] M. Muñoz-Aguirre, V. F. Ntasis, S. Rojas, and R. Guigó, "PyHIST: A histological image segmentation tool," *PLOS Comput. Biol.*, vol. 16, no. 10, Oct. 2020, Art. no. e1008349. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1008349
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778. [Online]. Available: https://ieeexplore.ieee.org/document/7780459/
[45] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015. [Online]. Available: http://link.springer.com/10.1007/s11263–015-0816-y
[46] F. Demir, A. Sengur, and V. Bajaj, "Convolutional neural networks based efficient approach for classification of lung diseases," *Health Informat. Sci. Syst.*, vol. 8, no. 1, Dec. 2020, Art. no. 4. [Online]. Available: http://link.springer.com/10.1007/s13755–019-0091-3
[47] R. Mohapatra, S. Saha, C. A. C. Coello, A. Bhattacharya, S. S. Dhavala, and S. Saha, "AdaSwarm: Augmenting gradient-based optimizers in deep learning with swarm intelligence," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 329–340, Apr. 2022.
[48] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, pp. 1–22, 2021.

Nikhilanand Arya received the BTech degree in information technology from the Maulana Abul Kalam Azad University of Technology, West Bengal, India, in 2015 and the MTech degree in mathematics and computing from the Indian Institute of Technology, Patna, India. He worked with Mindtree, India as a software engineer till 2017. He is currently working toward the PhD degree with the application of machine learning in bio-informatics, Indian Institute of Technology, Patna, India.

Archana Mathur received the master's degree in engineering from Bangalore University and the PhD degree is in machine learning and scientometrics. She is currently working as assistant professor with Nitte Meenakshi Institute of Technology. She has worked as research assistant with Indian Statistical Institute, Bangalore. Her area of interest lies in foundations of machine learning, deep learning and optimization techniques.

Snehanshu Saha (Senior Member, IEEE) received the PhD degreein mathematical sciences from the University of Texas at Arlington. He is a senior member of ACM and a fellow of IETE. He is a professor of Artificial Intelligence with BITS Pilani K K Birla Goa Campus. His current and future research interests lie in the theory of optimization, learning theory, activation functions in deep neural networks and astroInformatics.

Sriparna Saha received the PhD degree in computer science from Indian Statistical Institute Kolkata. She is currently associate professor with the Computer Science and Engineering Department of Indian Institute of Technology Patna, India. She has authored more than 300 technical papers and book chapters. Her publications currently report more than 4661 citations in Google Scholar (h-index:27). Her major research interests are evolutionary machine learning, deep learning, natural language processing and bioinformatics.