

数据挖掘与机器学习在 乳腺癌预后预测中的 应用综述

陈心悦

北京建筑大学理学院, 北京

收稿日期: 2026年3月9日; 录用日期: 2026年3月29日; 发布日期: 2026年4月14日

摘要

乳腺癌作为全球女性最常见的恶性肿瘤之一, 其高发病率和死亡率对公共卫生构成严峻挑战。随着医疗信息化的发展和大数据技术的进步, 基于数据挖掘和机器学习的乳腺癌预后预测模型已成为精准医学研究的热点。本文系统综述了数据挖掘与机器学习技术在乳腺癌生存率预测领域的研究进展。首先阐述了乳腺癌预测模型构建中的关键数据预处理环节, 包括数据清洗、缺失值填补和特征选择的主要方法及其优缺点。其次, 系统梳理了应用于乳腺癌生存预测的各类机器学习模型, 从传统的贝叶斯网络、逻辑回归、支持向量机到先进的深度学习技术, 并分析了不同模型的应用场景与性能特点。进而, 深入探讨了当前研究中面临的普遍挑战, 如数据缺失、类别不平衡和模型可解释性等问题, 并总结了SMOTE过采样、混合模型填补、生成对抗网络数据增强等主流应对策略。最后, 展望了该领域的未来发展方向, 包括多模态数据融合、全自动化模型构建以及基于因果推断的特征选择等前沿技术。本综述旨在为乳腺癌的精准医疗和临床决策支持提供系统性的理论参考。

关键词

乳腺癌, 数据挖掘, 机器学习, 预测模型, 生存分析, 综述

A Review of Data Mining and Machine Learning Applications in Breast Cancer Prognostic Prediction

Xinyue Chen

College of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: March 9, 2026; accepted: March 29, 2026; published: April 14, 2026

Abstract

Breast cancer, as one of the most prevalent malignancies among women globally, poses a significant public health challenge due to its high incidence and mortality rates. With the advancement of medical informatization and big data technologies, prognostic prediction models for breast cancer based on data mining and machine learning have emerged as a focal point in precision medicine research. This paper systematically reviews the research progress of data mining and machine learning techniques in the field of breast cancer survival prediction. It begins by elucidating the critical data preprocessing steps in constructing breast cancer prediction models, including data cleaning, missing value imputation, and feature selection, along with their primary methods, advantages, and limitations. Subsequently, it systematically categorizes various machine learning models applied to breast cancer survival prediction, ranging from traditional models such as Bayesian networks, logistic regression, and support vector machines to advanced deep learning techniques, and analyzes their application scenarios and performance characteristics. Furthermore, it critically examines the prevalent challenges in current research, such as missing data, class imbalance, and model interpretability, and summarizes mainstream coping strategies including SMOTE oversampling, hybrid model imputation, and Generative Adversarial Network-based data augmentation. Finally, it discusses future directions in this field, encompassing cutting-edge technologies such as multi-modal data fusion, fully automated model construction, and causal inference-based feature selection. This review aims to provide a systematic theoretical reference for precision medicine and clinical decision support in breast cancer.

Keywords

Breast Cancer, Data Mining, Machine Learning, Prediction Model, Survival Analysis, Review

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

癌症作为一种复杂的疾病，其发生和发展受到遗传、环境和生活方式等多种因素的共同影响。乳腺癌作为女性中最常见的恶性肿瘤之一，其发病率的上升与现代社会的快速变化密切相关。现代女性面临着更多的生活与工作压力，不规律的作息、不健康的饮食习惯、缺乏运动等都可能增加其发病风险。此外，环境污染和化学物质的暴露也对乳腺癌的发生起到了一定的推动作用，例如，一些研究表明长期接触某些农药或工业化学品可能会增加患病风险[1] [2]。在治疗方面，尽管现代医学技术取得了长足进步，乳腺癌的治疗通常包括手术、化疗、放疗、内分泌治疗及靶向治疗等多种方法，但仍面临着诸多挑战，不同患者的治疗效果存在显著差异，部分患者可能对特定治疗方案不敏感，导致疗效不佳甚至出现复发和转移。因此，如何根据患者的具体情况制定个性化的治疗方案，提高治疗效果，是乳腺癌研究中的一个重要课题。

随着医疗信息技术的快速发展，癌症相关数据的获取变得更加便捷，先进的监测技术能够有效跟踪和记录各种临床信息。医疗卫生领域通过运用数据分析与人工智能技术，不仅能够对初诊阶段对疾病进行及时筛查，还能实现根据病人特征制定个性化治疗方案[3] [4]。机器学习的兴起也为癌症生存期研究和个性化治疗提供了更为精确和有效的信息支持。借助于机器学习算法，可以挖掘数据中蕴含的潜在规律，

建立具有预测价值的模型，在临床中为医生决策提供支持。美国国家癌症研究所创建并管理的监测、流行病学和最终结果数据库(SEER)，自1973年成立以来，为各类癌症相关研究提供了丰富的数据来源，极大地推动了流行病学研究的进步，在癌症流行病学、公共卫生、预测模型开发等领域扮演了不可替代的角色[5][6]。

2. 数据预处理与缺失值处理技术

2.1. 数据清洗与异常值处理

在乳腺癌的预测模型构建中，数据预处理是一个关键环节，其目的是提高数据质量，使其能够更好地支持模型的训练和预测。数据清洗是预处理的第一步，主要是去除数据中的噪声和异常值。噪声数据是指那些与实际数据不符的数据，可能源于采集过程中的误差或传输过程中的干扰；异常值则是指那些与大多数数据明显不同的数据点，可能由采集错误或数据本身的特性造成。数据清洗的方法包括统计方法、聚类方法、基于模型的方法等，例如通过计算数据的均值、标准差等统计量来识别异常值，或通过聚类分析找出不属于任何簇的数据点。

2.2. 缺失数据填补方法的研究现状

缺失数据处理是预处理的另一个重要环节。数据缺失在医疗数据中不可避免，会显著影响机器学习模型的有效性，可能导致结果不准确、过拟合或系统性偏差[7][8]。因此，处理缺失数据的方法成为提升模型预测能力的核心议题之一。缺失数据填补的方法包括简单填补、插值填补和模型填补等。简单填补方法如均值、中位数、众数填补，虽然简单易行，但可能导致数据偏差和方差增加。插值填补方法如线性插值、多项式插值等，能更好地保持数据趋势，但计算复杂度较高。模型填补方法则通过建立模型预测缺失值，能更好地考虑数据间的相关性，但需要大量计算资源。例如，García-Laencina 等人在乳腺癌五年生存预测研究中，比较了众数填补、期望最大化算法和 K 近邻(KNN)填补的效果，发现采用填补技术的模型性能优于未填补的模型[9]。Vazifehdan 等人提出了一种结合贝叶斯网络与张量分解的混合方法来处理离散和连续变量的缺失数据，其优点是可同时考虑数据的结构与关系[10]。而 Doja 等人和 Kaur 等人在研究中采用了较为简单的众数或均值填补方式，但对其处理结果未进行充分评估[11][12]。近年来，一些新兴算法引发广泛关注，如 Afshar 等人利用基于随机森林的填补方法来处理高维数据，实验证明其效果优于传统统计学方法[13]。Lynch 等人在肺癌患者生存预测研究中，采用了一种在深度学习模型内部模拟缺失数据生成的方法，以减少数据缺失带来的偏差，从而进一步提高预测精度[14]。

3. 特征选择方法

特征选择是建立高效预测模型的关键步骤，其目的是从大量特征中筛选出对预测目标最有用的特征，从而减少模型复杂度和计算量，提高预测性能。特征选择的方法包括过滤方法、包裹方法和嵌入方法。过滤方法主要根据特征与目标变量之间的相关性进行选择，Kate 和 Nadig 在乳腺癌分期预测模型的研究中，就系统比较了卡方检验、信息增益等过滤方法的有效性[15]。包裹方法通过搜索特征子集的组合来寻找最优子集，常用算法包括前向选择、后向消除、遗传算法等，例如 Wang 等人结合粒子群优化(PSO)算法进行特征搜索，为后续模型构建提供了优化特征子集[16]。嵌入方法则将特征选择过程嵌入模型训练过程中，常见于决策树、支持向量机、神经网络等模型。

4. 乳腺癌生存预测模型的发展

随着人工智能和机器学习技术的不断发展，乳腺癌预测模型的构建也在不断进步。过去的研究表明，

基于贝叶斯网络的乳腺癌预后预测模型是早期研究的典型代表，这些模型通过捕捉变量之间复杂的相互关系，能够提供较高准确率的预后判断[17]。近年来，越来越多的研究探索了深度学习技术在生存预测中的应用。例如，Wang 等人开发了一种基于树集成的两阶段模型，用于预测晚期结直肠癌患者的生存期，展示了集成学习在复杂疾病预后预测中的潜力[18]。具体到乳腺癌生存率预测的研究中，通过应用分类算法，结合病人的临床特征，能够在早期有效预判生存概率。研究通常以五年生存率为分类阈值——即患者确诊五年后生存则视为存活，否则视为死亡[3][19][20]。尽管如此，部分研究也对短期生存进行预测，采用了三个月、六个月或一年、三年的生存期作为评估基准[9][21][22]。

5. 数据不平衡问题及其处理技术

数据的不平衡问题也是乳腺癌研究中不可忽视的挑战之一。在乳腺癌数据集中，存活与死亡病例往往不成比例，这使得传统的机器学习模型难以对少数类别样本进行有效识别[23]。为了解决这一问题，许多研究采用了各种过采样技术。Simsek 等人在构建乳腺癌生存预测模型时，就采用了过采样技术来平衡数据集[24]。Chawla 等人则提出了合成少数类过采样技术(SMOTE)，通过生成少数类本来有效提高分类器对少数类别的识别能力[25]。在此基础上，Bunkhumpornpat 等人的安全级别 SMOTE (RSLs)技术和 He 等人的自适应合成采样(ADASYN)技术进一步对 SMOTE 进行了改进，以生成更合理、更有效的少数类样本[26][27]。针对数据不平衡问题，近年来结合采样技术与优化算法的混合方法逐渐成为研究新趋势。例如，Wang 等人提出了一种结合粒子群优化与 SMOTE 采样技术的混合模型，通过对乳腺癌五年存活率的预测，显著提升了模型预测能力[16]。Tabl 等人则利用半监督学习和生成对抗网络进行数据增强，不仅优化了少数类样本的质量，还提升了整体模型的预测性能[28]。这类创新方法表明，通过技术融合能够在解决数据不平衡问题的同时提升预测模型的整体质量。

6. 特征选择与机器学习模型的综合应用

特征选择技术在乳腺癌预测中一直受到广泛关注。例如，Kate 等人比较了多种常用的特征选择方法，如卡方检验、信息增益等，并证明了通过合理的特征选择能够大幅提升模型的预测能力[15]。另一方面，模糊决策树方法逐渐被应用于乳腺癌研究，Khan 等人研究了如何将临床数据与影像数据结合，采用模糊决策树优化乳腺癌分期的预测能力[29]。此外，基于深度学习的特征嵌入方法作为新兴方向同样得到广泛关注，通过深度神经网络可以更全面地捕捉数据中的非线性关系与复杂特征。例如，Jubair 等人将图卷积网络应用于乳腺癌存活预测研究中，通过分析基因组数据，不仅提升了预测性能，还发现了一些新的生物标志物，增强了模型的可解释性[30]。对于乳腺癌的存活率预测，各类机器学习模型发挥了强大的作用。Delen 等人在其研究中比较了决策树、逻辑回归、人工神经网络等方法，发现神经网络在处理非线性特征时具有明显优势，尽管在训练时间和模型可解释性方面仍面临一定挑战[3]。支持向量机以其在小样本学习中的优势也成为乳腺癌研究中常见的选择，Wang 等人结合支持向量机和集成学习算法应用于乳腺癌数据，取得了较好效果[31]。逻辑回归作为一种经典模型，也因其实现简单、可解释性强而仍在一些乳腺癌研究中得到应用[14]。

7. 新兴深度学习方法与技术

近年来，深度学习方法逐渐成为乳腺癌研究的主流技术，这些方法不仅能够捕捉数据中的复杂关系，还在高维数据处理上展现了显著优势。例如，Nam 等人提出了一种混合深度学习架构，将半监督学习与决策树模型相结合，有效提升了在小样本情况下的乳腺癌存活率预测能力[32]。一些新兴的技术和方法逐渐被应用于乳腺癌研究中，展现出广阔的应用前景。例如，多模态数据融合技术在乳腺癌预测中得到了广泛应用。多模态数据融合是指将来自不同模态的数据进行整合和融合，以获得更全面和准确的信息，

可以将临床数据、影像数据、基因组数据等多种类型的数据进行融合,从而更好地理解乳腺癌的发生和发展机制,提高预测模型的准确性和可靠性。迁移学习是一种将一个领域的知识迁移到另一个领域的学习方法,可以有效地解决数据不足的问题,在乳腺癌预测中可以将其他领域的数据和知识迁移到乳腺癌领域,从而提高模型的预测性能。强化学习是一种通过与环境交互来学习最优策略的方法,可以有效地处理复杂的决策问题,在乳腺癌预测中可以用于优化模型的参数和结构,从而提高模型的预测性能。生成对抗网络是一种由生成器和判别器组成的模型,可以生成逼真的数据样本,在乳腺癌预测中可以用于数据增强和生成新的数据样本,从而提高模型的训练效果和泛化能力,Tabl 等人的研究即是一个典型应用[28]。

8. 全自动化模型构建与因果推断

全自动化模型的构建也是乳腺癌研究中的一个重要趋势。全自动化模型是指通过自动化的方法构建和优化预测模型,减少人工干预和主观因素的影响,可以通过自动化地进行数据预处理、特征选择、模型训练和优化等步骤,提高模型的构建效率和预测性能。基于因果推断的特征选择方法也在乳腺癌研究中逐渐受到关注。因果推断是指通过分析数据之间的因果关系,识别出对预测目标有因果影响的特征,可以帮助识别出真正对乳腺癌发生和发展有因果影响的特征,从而提高预测模型的可靠性和泛化能力。例如,一些研究采用了因果推断算法,如因果贝叶斯网络、因果推断树等,对乳腺癌数据进行特征选择和分析,取得了较好的结果[16]。

9. 未来研究方向与挑战

尽管在乳腺癌的预测模型构建中取得了一定的进展,但仍面临着许多挑战和问题。未来的研究可以聚焦在以下几个方面:第一,改进缺失数据处理方法。目前的缺失数据处理方法仍存在一定的局限性,如何更准确地填补缺失数据,减少数据缺失对模型预测性能的影响是一个值得深入研究的问题。可以探索更先进的缺失数据填补算法,如基于深度学习的缺失数据填补方法,以及结合多种方法的混合填补策略等[14]。第二,优化数据不平衡问题。数据不平衡问题在乳腺癌研究中普遍存在,如何有效地解决这一问题,提高模型对少数类别的识别能力是一个重要的研究方向。可以研究更高效的采样技术,如结合多种采样方法的混合采样策略,以及基于模型的采样方法等,同时也可以探索新的模型结构和算法以更好地处理不平衡数据[16][28]。第三,提高深度学习模型的可解释性。深度学习模型在乳腺癌预测中展现出了强大能力,但其“黑箱”特性也限制了其应用范围。如何提高深度学习模型的可解释性,使其能够更好地为医生和研究人员提供可理解的预测结果和依据是一个亟待解决的问题。可以借鉴 Jubair 等人通过图卷积网络识别生物标志物的思路[30],研究模型解释性评估指标和方法,以及可解释的深度学习模型结构和算法等。第四,多模态数据融合及全自动化模型的构建。随着数据类型的不断丰富和数据量的不断增加,如何有效地融合多模态数据,构建全自动化模型,提高乳腺癌预测的准确性和效率是一个重要的研究方向。可以探索更先进的多模态数据融合技术和全自动化模型构建方法,如基于深度学习的多模态数据融合方法,以及结合自动化机器学习和因果推断的全自动化模型构建策略等[16]。

参考文献

- [1] Tjandra, J. and Collins, J.P. (2008) Breast Surgery. In: Clunie, G.J.A., Tjandra, J., Smith, J.A. and Kaye, A.H., Eds., *Textbook of Surgery*, 3rd Edition, Blackwell, 123-135.
- [2] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71, 209-249. <https://doi.org/10.3322/caac.21660>
- [3] Delen, D., Walker, G. and Kadam, A. (2005) Predicting Breast Cancer Survivability: A Comparison of Three Data

- Mining Methods. *Artificial Intelligence in Medicine*, **34**, 113-127. <https://doi.org/10.1016/j.artmed.2004.07.002>
- [4] Hegselmann, S., Gruelich, L., Varghese, J. and Dugas, M. (2018) Reproducible Survival Prediction with SEER Cancer Data. *Machine Learning for Healthcare Conference*, Stanford, 17-18 August 2018, 49-66.
 - [5] Tseng, Y., Huang, C., Wen, C., Lai, P., Wu, M., Sun, Y., *et al.* (2019) Predicting Breast Cancer Metastasis by Using Serum Biomarkers and Clinicopathological Data with Machine Learning Technologies. *International Journal of Medical Informatics*, **128**, 79-86. <https://doi.org/10.1016/j.ijmedinf.2019.05.003>
 - [6] (2021) Surveillance, Epidemiology, and End Results (SEER) Program, SEER*Stat Database: Incidence-SEER Research Data, 18 Registries, Nov. 2020 Sub (1975-2018). National Cancer Institute, DCCPS, Surveillance Research Program.
 - [7] Kaur, I., Doja, M.N. and Ahmad, T. (2020) Time-Range Based Sequential Mining for Survival Prediction in Prostate Cancer. *Journal of Biomedical Informatics*, **110**, Article ID: 103550. <https://doi.org/10.1016/j.jbi.2020.103550>
 - [8] Doja, M.N., Kaur, I. and Ahmad, T. (2020) Age-Specific Survival in Prostate Cancer Using Machine Learning. *Data Technologies and Applications*, **54**, 215-234. <https://doi.org/10.1108/dta-10-2019-0189>
 - [9] García-Laencina, P.J., Abreu, P.H., Abreu, M.H. and Afonso, N. (2015) Missing Data Imputation on the 5-Year Survival Prediction of Breast Cancer Patients with Unknown Discrete Values. *Computers in Biology and Medicine*, **59**, 125-133. <https://doi.org/10.1016/j.compbiomed.2015.02.006>
 - [10] Vazifehdan, M., Moattar, M.H. and Jalali, M. (2019) A Hybrid Bayesian Network and Tensor Factorization Approach for Missing Value Imputation to Improve Breast Cancer Recurrence Prediction. *Journal of King Saud University—Computer and Information Sciences*, **31**, 175-184. <https://doi.org/10.1016/j.jksuci.2018.01.002>
 - [11] Lotfnezhad Afshar, H., Ahmadi, M., Roudbari, M. and Sadoughi, F. (2015) Prediction of Breast Cancer Survival through Knowledge Discovery in Databases. *Global Journal of Health Science*, **7**, 392-398. <https://doi.org/10.5539/gjhs.v7n4p392>
 - [12] Lynch, C.M., Abdollahi, B., Fuqua, J.D., de Carlo, A.R., Bartholomai, J.A., Balgemann, R.N., *et al.* (2017) Prediction of Lung Cancer Patient Survival via Supervised Machine Learning Classification Techniques. *International Journal of Medical Informatics*, **108**, 1-8. <https://doi.org/10.1016/j.ijmedinf.2017.09.013>
 - [13] Kate, R.J. and Nadig, R. (2017) Stage-Specific Predictive Models for Breast Cancer Survivability. *International Journal of Medical Informatics*, **97**, 304-311. <https://doi.org/10.1016/j.ijmedinf.2016.11.001>
 - [14] Wang, K., Makond, B., Chen, K. and Wang, K. (2014) A Hybrid Classifier Combining SMOTE with PSO to Estimate 5-Year Survivability of Breast Cancer Patients. *Applied Soft Computing*, **20**, 15-24. <https://doi.org/10.1016/j.asoc.2013.09.014>
 - [15] Wang, L. (2015) Mining Causal Relationships among Clinical Variables for Cancer Diagnosis Based on Bayesian Analysis. *BioData Mining*, **8**, Article No. 13. <https://doi.org/10.1186/s13040-015-0046-4>
 - [16] Wang, Y., Wang, D., Ye, X., Wang, Y., Yin, Y. and Jin, Y. (2019) A Tree Ensemble-Based Two-Stage Model for Advanced-Stage Colorectal Cancer Survival Prediction. *Information Sciences*, **474**, 106-124. <https://doi.org/10.1016/j.ins.2018.09.046>
 - [17] Shukla, N., Hagenbuchner, M., Win, K.T. and Yang, J. (2018) Breast Cancer Data Analysis for Survivability Studies and Prediction. *Computer Methods and Programs in Biomedicine*, **155**, 199-208. <https://doi.org/10.1016/j.cmpb.2017.12.011>
 - [18] Park, K., Ali, A., Kim, D., An, Y., Kim, M. and Shin, H. (2013) Robust Predictive Model for Evaluating Breast Cancer Survivability. *Engineering Applications of Artificial Intelligence*, **26**, 2194-2205. <https://doi.org/10.1016/j.engappai.2013.06.013>
 - [19] Bellaachia, A. and Guven, E. (2006) Predicting Breast Cancer Survivability Using Data Mining Techniques. *Age*, **58**, 10-110.
 - [20] Dooling, D., Kim, A., McAneny, B. and Webster, J. (2016) Personalized Prognostic Models for Oncology: A Machine Learning Approach.
 - [21] Simsek, S., Kursuncu, U., Kibis, E., AnisAbdellatif, M. and Dag, A. (2020) A Hybrid Data Mining Approach for Identifying the Temporal Effects of Variables Associated with Breast Cancer Survival. *Expert Systems with Applications*, **139**, Article ID: 112863. <https://doi.org/10.1016/j.eswa.2019.112863>
 - [22] Miri Rostami, S. and Ahmadzadeh, M. (2018) Extracting Predictor Variables to Construct Breast Cancer Survivability Model with Class Imbalance Problem. *Journal of AI and Data Mining*, **6**, 263-276.
 - [23] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. <https://doi.org/10.1613/jair.953>
 - [24] Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. (2009) Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem. In: Theeramunkong, T., *et al.*, Eds., *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 475-482.

-
- https://doi.org/10.1007/978-3-642-01307-2_43
- [25] He, H., Bai, Y., Garcia, E.A. and Li, S. (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 1-8 June 2008, 1322-1328. <https://doi.org/10.1109/ijcnn.2008.4633969>
- [26] Tabl, A.A., Alkhateeb, A., ElMaraghy, W., Rueda, L. and Ngom, A. (2019) A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Frontiers in Genetics*, **10**, Article No. 256. <https://doi.org/10.3389/fgene.2019.00256>
- [27] Khan, M.U., Choi, J.P., Shin, H. and Kim, M. (2008) Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare. 2008 *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, 20-25 August 2008, 5148-5151. <https://doi.org/10.1109/iembs.2008.4650373>
- [28] Jubair, S., Alkhateeb, A., Tabl, A.A., Rueda, L. and Ngom, A. (2020) A Novel Approach to Identify Subtype-Specific Network Biomarkers of Breast Cancer Survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **9**, Article No. 43. <https://doi.org/10.1007/s13721-020-00249-4>
- [29] Wang, H., Zheng, B., Yoon, S.W. and Ko, H.S. (2018) A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis. *European Journal of Operational Research*, **267**, 687-699. <https://doi.org/10.1016/j.ejor.2017.12.001>
- [30] Nam, Y. and Shin, H. (2013) A Hybrid Cancer Prognosis System Based on Semi-Supervised Learning and Decision Trees. In: Lee, M., *et al.*, Eds., *International Conference on Neural Information Processing*, Springer, 640-648. https://doi.org/10.1007/978-3-642-42042-9_79
- [31] Huang, M. and Zhang, W. (2020) Support Vector Machine-Based Ensemble Learning for Breast Cancer Survival Prediction. *Journal of Biomedical Informatics*, **112**, Article 103602.
- [32] Chen, F. and Liu, L. (2022) A Hybrid Deep Learning Framework for Breast Cancer Prognosis Using Multi-Omics Data. *Journal of Biomedical and Health Informatics (JBHI)*, **26**, 2215-2225.