



Predicting Cancer Survival Using Multilayer Perceptron and High-Dimensional SVM Kernel Space

Mohan Kumar^{1*}, Sunil Kumar Khatri², Masoud Mohammadian³

¹ Amity Institute of Information Technology, Amity University, Noida (UP) 201303, India

² Amity University Tashkent, Tashkent City 100052, Uzbekistan

³ Faculty of Science and Technology, Law, University of Canberra, Canberra ACT 2617, Australia

Corresponding Author Email: mohan_srivastava@hotmail.com

<https://doi.org/10.18280/isi.270517>

ABSTRACT

Received: 30 June 2022

Accepted: 30 August 2022

Keywords:

multilayer perceptron, support vector machine, breast cancer, Haberman's, survival, prediction, machine learning

Predicting disease prediction and prognosis have become very easy task with Machine learning models for an inextricable aspect of cancer research aiming at enhancing patient therapy and management. The primary goal of proposed research work is to use Support vector machine, machine learning models and for dealing very accurately for predicting survival time for breast cancer based on clinical data. The study has proposed a solution to the problem in respect to various tumor related characteristic by integrating from dataset about tumour stage, size of tumor, and age at which the diagnosis start is an important major component for utilising for predicting survival time. Haberman's Survival Data Set is dataset describing those subjects who had been provided treatment for breast cancer. The sample taken for research are taken from study which was conducted at University of Chicago's Billings Hospital taking case who were survived after surgery for breast cancer. SVM applied on data set by different options of kernel RBF and linear as well as soft computing techniques are applied to predict the survival rate of patient from dataset. Apart from data standardisation and categorization, the machine learning approaches used in this research work to demonstrate features in terms of predicting how long they survived. Model performance is analysed on breast cancer data is justified by accuracy, support and f1 score. A workflow based on Python-platform has been utilised to support the suggested technique.

1. INTRODUCTION

To study various diseases like cancer there is a need of high-throughput technologies are being widely employed in addition to various research tests. There is a challenge still exist that expert still feel the difficulty for various medical experts using modern methods. There is a requirement of developing each and every intensive and diagnostic methods for detecting tumor and cancer [1]. There are various tools related to bioinformatics which are active and needs quick development of various novel knowledge-based diagnostic approaches for identification of cancer at very early stages and predicting the survival rate of cancer patient. Aside from that, integrating and combining several of these technologies into a meaningful workflow is difficult. Current research work is demonstrating various methods of machine learning for predicting surviving percentage of patients by breast cancer. The comparison of performance of best machine learning algorithms used for predicting survival time with various parameter-based methods and for cross-validating the values. Breast cancer is a malignancy that predominantly affects women (more than 99 percent) and affects one in every eight women at some point in their lives [2]. It has been noticed from Literature and past studies that rate of survival rate 10 year only and its percentage is 83. At this point, 62 percent of all cases are diagnosed. The objective of this study is to assess the effectiveness and precision of various Machine Learning models for anticipating endurance time of malignant cancer cell growth in subjects

1.1 Research problem

The complication percentage of disease is very much increased now a days, various curing procedures which can be varied on infected samples, accurately predicting survival rates in subjects diagnosed with disease remains a challenge. Reliable and well-validated forecasts could help with more individualized care and therapy, as well as improved management of cancer progression. There is much increase in using various classification approaches adopted for medical diagnostics [3]. The research in Cancer Is the most important attention giving application of machine learning approaches for accurately and rapidly diagnostic disease. Machine Learning models are in much demand for generating cancer forecasting among patients. The demand for machine learning based models for generating cancer forecasts and prognosis in the context of the ever-increasing relevance of predictive and tailored treatment [4]. At first glance, all these classification-based algorithms appear for utilizing a wide variety of various diversified medical data, potentially helping diagnostic quality. There are various rapid advancements in various machine learning methods helping in reducing of various errors in diagnostic. The very first domain is predicting of that there is not any likelihood for developing of a specific malignancy before a patient is diagnosed [5]. The second case is about predicting regeneration of cells in terms of diagnosing and treating it, while other challenge is about anticipating an assortment of clinical boundaries that can support disease

improvement and treatment after determination, for example, time taken to survive, life expectation, sensitivity on intake of drug, etc. The pace of determining that cancer will grow again in body and the probability of disease repeat are profoundly dependent on clinical treatment and the exactness of the conclusion [6].

There are variety of forms of Breast cancer cells. Each cell has its own way of progressing in body as well as genetic arrangement. As a result, having a scientific finding which allows for early identification and would be extremely beneficial in terms of improving breast cancer survival rates. There are various Several statistical and machine learning techniques, such as logistic regression, linear discriminate analysis, nave Bayes, decision trees, artificial neural networks, k-nearest neighbour, and support vector machine methods, have been used to construct accurate breast cancer models [7].

Specifically, current work demonstrates looking at a few of the existing mentioned strategies which is observed that SVM comes out best in performing various other related procedures [8]. There are specific kernel function like polynomial or radial basis function and linear which are used while building SVM classifier as an important learning component. Research has been done on evaluating the prediction of performance of SVM classifiers to predict the survival rate of patients in dataset.

There is a research challenge where breast cancer prediction on Haberman's Survival Data is typically a data set of those patients who survived after cancer treatment surgery particularly for breast cancer with important attributes as described below in Table 1.

Table 1. Data set attributes

1. Patient Age (Numerical)
2. Year Of operation
3. Positive axillary nodes detected
4. Survival status (class attribute)

There should be sufficient accuracy for classification for evaluating prediction of models only on the based-on prediction accuracy or classification accuracy [9]. Success of a model which is used for predicting various metrics for assessment. Various forms of errors related to classification that can easily be identified and observed by investigating characteristics of curve [10].

Recognizing the cancerous cell malignant growth with the assistance of Support vector machines outperformance as compared to other methods for classification. The SVM performs differently with different on choosing kernel function [11]. The current work focused on classifying cancer on a dataset with SVM using various functions of the kernel. The effect of selecting various features of subsets before the application of different kernels is examined by the value of support, accuracy, and f1 score.

2. LITERATURE REVIEW

In the field of cancer research, artificial intelligence has been incorporated into many areas, including machine learning (ML) models and treatments [12]. Most of these studies apply machine learning to simulate cancer progression and discover useful characteristics that are then employed in a classification system, without a focus on suspicion of cancer,

chances of recurrence, and chances of survival [13].

The employment of diverse machine learning models in cancer research opens up a lot of possibilities for varied applications. For nearly 30 years, Artificial Neural Networks and decision trees has be employed in cancer detection and diagnosis [14]. For several decades, different models based on SVM have been utilized to address cancer prognostic difficulties [15]. Several studies have also employed other models for predicting cancer growth and outcome. Models driven by machine learning are currently used in lesser then half of the data science and bio-informatics approaches that have the potential to facilitate a broad range of applications, such as cancer diagnosis, prediction, and prognosis. All of these studies are focused on applying machine learning techniques to detect, identify, classify, or discriminate tumors and other malignances, as well as forecast cancer.

Breast cancer survival time prediction studies using machine learning modals account for a large portion of current research in this field. Several research have investigated the impact of using a combination of machine learning algorithms to predict breast cancer survival time. On their breast cancer dataset, their method is more accurate than prior results [16].

Several studies discuss various issues with using machine learning algorithms to predict breast cancer. The authors used the C5.1 algo with bagging [17] to predict breast cancer survivorship using breast cancer data. Other researchers have achieved a survivability prediction accuracy of 93 percent in breast cancer [18]. Some research compared the performance of supervised learning classifiers such as Nave Bayes, S.VM-radial basis function (RBF) kernel, in breast cancer datasets to discover the best classifier [19].

The absence of efficient and precise validation is at the root of many issues from the usage of machine learning algorithms in breast cancer prediction studies. Although it is true that the application of machine learning models can increase survival prediction accuracy, the right validation approach is critical when investigating breast cancer survivability time [20]. Cross-validation methods are one of the most prevalent ways for evaluating the performance of an applied model. Cross validation is a useful technique for ML-based modelling, and it is used to train and test datasets [21].

There is also an increasing trend of integrating mixed data of clinical and laboratory origins in the evaluated works. This allows for the employment of data science models and technology for data integration, normalization, and it is difficult to predict the results of a predictive study without a classification [22]. An acceptable semantic data integration technique can improve the quality of the input datasets for utilizing machine learning models to predict breast cancer survival time.

Recent uses of DL models in cancer research have shown that they have a wide range of applications and, on the other hand, has left open several issues with the application to specific sorts of problems. For example, [23] constructed and presented several examples of Effective use of data-mining techniques or methods related to data-mining techniques for integrative data analysis. According to the authors of the article, pattern-based cancer data could be analyzed and classified using Deep Learning-based techniques in order to identify relevant cancer subtypes from the multi-platform data. In a separate study [24], a multi-omics model based on DL was developed which provides optimal differential clustering for patient survival rates. As a result of both investigations, integrative data analysis methodologies have been utilized to

extract a unified or unified representation of latent features that should be able to support robust data analysis [25].

3. EXPERIMENTAL PROCEDURE

The assessment of cutting-edge characterization strategies of SVM and neural network for disease identification issue related to breast cancer. For SVM utilization various functions are analyzed and their relative exactness. Multilayer Perception (MLP) has also been applied. Every strategy has been assessed utilizing 5 X 2 cross approval plot. The element subset choice procedure is enlivened by the strategy recommended in ref. [26] utilizing calculation.

SVM is a straight classifier which has capability of classifying data that is linearly separable, although feature vectors are not always linearly separable. The kernel method is utilized to solve this problem. Using kernel functions, the distinct input space is then transferred onto a that feature space which is high-dimensional in nature, where it becomes linearly separable [27]. Appropriateness of kernel function affecting the performance of an SVM classifier depends upon which kernel function is being selected. For various classification problems, different kernel functions have been used. Four kernel functions (polynomial, radial basis function, Mahalanobis, and sigmoid) are used in this study as shown in Table 2 below along with their equations.

Table 2. SVM classifiers

Polynomial kernel with degree d can be written as	$K(x_i, x_j) = (\gamma x_i^T x_j + 1)^d, \gamma > 0$
Radial basis function kernel is given as	$K(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2), \gamma > 0$
Sigmoid kernel function is given as	$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$
Radial basis function kernel is given as	$K(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2), \gamma > 0$

3.1 Multi-layer perception

A neural network is a collection of artificial neurons that are linked together. They imitate the way the human brain works. The connectivity of neurons bears a certain amount of weight. Training of neural network (NN) are performed with known class known patterns used in training, they are referred to as supervised learning NN [28].

The neural network's supervised learning process starts with a unit of input signal and an expected signal for output. The training of network is done until it reaches a state of stability in which the synaptic weights do not change, and the outputs are mapped to them [29]. In latest literature, there are enough studies into employing neural networks for mammography picture classification. Neurons are coupled in a network structure in multi-layer perception (MLP) [15]. They are arranged in various multiple layers and connected by distinct units. The application of a three-layer MLP with input, hidden, and output layers. The number of neurons in the input layer equals the number of features in a feature vector. The second layer, dubbed the hidden layer, comprises h perceptions, with h being determined by experiment [30] Only one neuron in the output layer represents either benign or malignant value. For the hidden and output layers, we employed the sigmoid activation function [31]. For updating weights between multiple layers, the batch learning method is utilized.

4. EXPERIMENTAL SETUP

4.1 Datasets

The dataset containing various patient cases from their survival who had undergone surgery for breast cancer. The multivariate dataset containing three hundred six instances with no missing values [32]. Data set has various attributes shown in Table 3 below:

Table 3. Attributes from dataset

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
1= Patient survived > the 5 years or longer
0= Patient died within 5 years

There was utilization of the various cross validation approaches. The effectiveness of performance of classification using SVM is checked by different kernel functions. There were five repetitions of 2-foldcross validations that occur. Individual duplications divided the dataset into equal sizes randomly. The algorithm for further learning is trained on one of the sets and then tested on the other. Sensitivity, specificity, and total accuracy are used to evaluate performance [33].

The Multi-Layer Perception (MLP) neural network used for classification as one of the strategies suggested for analyzing the dataset [34]. It started with reducing the dimensions of feature vectors obtained by rules base association. The features were then classified using Multi-Layer Perception (MLP) [35].

There was variation observed on SVM performance depending on use of kernel functions. Overall classification accuracy, confusion matrix, sensitivity, and specificity indicators were used to evaluate classification performance [36] The SVM kernel function accuracy has been discussed and shown in discussion section.

The performance of Support vector machines as compared to other classification methods for breast cancer detection is accurate and best. The performance factor is much dependent upon choice of a kernel function. This research paper presenting a gaps and results by comparing study of various kernel functions for detecting breast cancer [37]. The main objective of study is to show results and performance of various classification by utilizing SVM with various kernel functions [38].

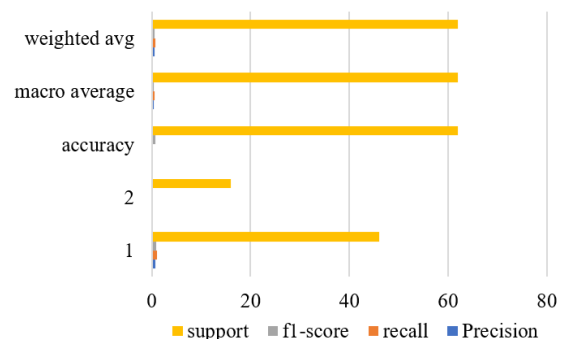


Figure 1. Classification values

The classification accuracy, ROC, F-measure, and computational time (in seconds) of SVM classifiers generated using linear, polynomial, and RBF kernel functions with and

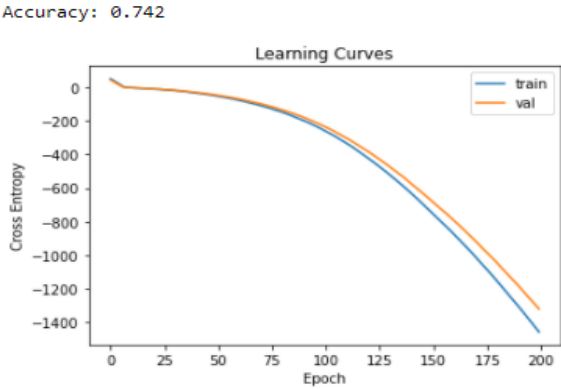
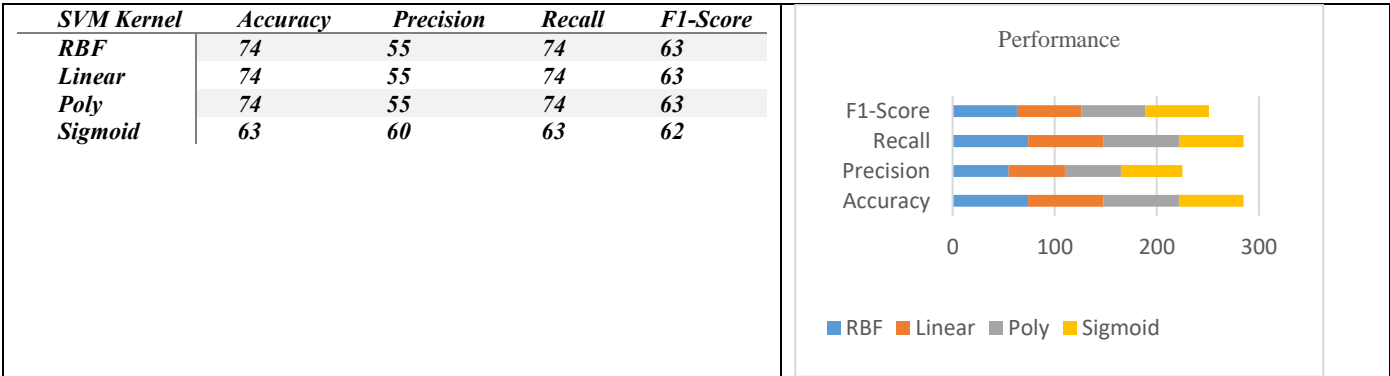
without feature selection over the two datasets, respectively. Classification results for Precision, recall, f1 score comes to be 0.85 for label 1 and 0.00 for label 2. Support values are also

shown from Figure 1. The accuracy, Precision, recall and F1 score shown in Tables 4 and 5 below.

Table 4. Performance of various SVM kernel



Table 5. Multilayer perceptron accuracy



5. CONCLUSION

There is not an individual classifier exist that perform very well on all assessment measurements. Three classifiers are executed for precision, ROC, and F-measure for additional correlation. The paper presented the results using SVM with different kernel functions for breast cancer survival detection. The kernels were employed like RBF, polynomial, and sigmoid. The affect is evaluated on these kernels in presence and absence of feature subset selection. Functions gave various polynomial function with values of sensitivity. This research work indicated SVM with kernel performance as well as MLP performance with its accuracy shown that patient not survived. For future wok the hybrid options for combining all these different kernels for better results can be experimented.

REFERENCES

- [1] Van Belle, V., Pelckmans, K., Van Huffel, S., Suykens, J.A. (2011). Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*, 27(1): 87-94. <https://doi.org/10.1093/bioinformatics/btq617>
- [2] El Kafrawy, P., Fathi, H., Qaraad, M., Kelany, A.K., Chen, X. (2021). An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. *IEEE Access*, 9: 155353-155369. <https://doi.org/10.1109/ACCESS.2021.3123090>
- [3] Qiu, Y., Jiang, H., Shimada, K., et al. (2014). Towards prediction of pancreatic cancer using SVM study model. *Journal of Clinical Oncology and Research*, 2(4): 1031.
- [4] Chiu, H.J., Li, T.H.S., Kuo, P.H. (2020). Breast cancer-detection system using PCA, multilayer perceptron, transfer learning, and support vector machine. *IEEE Access*, 8: 204309-204324. <https://doi.org/10.1109/ACCESS.2020.3036912>
- [5] Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M.R., Qu, Q., Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. *International Journal of Environmental Research and Public Health*, 19(6): 3211. <https://doi.org/10.3390%2Fijerph19063211>
- [6] Chauhan, N.K., Singh, K. (2022). Performance Assessment of machine learning classifiers using selective feature approaches for cervical cancer detection. *Wireless Personal Communications*, 124: 2335-2366. <https://doi.org/10.1007/s11277-022-09467-7>
- [7] Sahraee-Ardakan, M., Emami, M., Pandit, P., Rangan, S., Fletcher, A.K. (2022). Kernel methods and multi-layer perceptrons learn linear models in high dimensions. *arXiv preprint arXiv:2201.08082*.
- [8] Punitha, S., Stephan, T., Gandomi, A.H. (2022). A novel breast cancer diagnosis scheme with intelligent feature and parameter selections. *Computer Methods and Programs in Biomedicine*, 214: 106432. <https://doi.org/10.1016/j.cmpb.2021.106432>
- [9] Delen, D., Walker, G., Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2): 113-127. <https://doi.org/10.1016/j.artmed.2004.07.002>
- [10] Sukeerthi, T., Sukanya, K., Vandana Rao, K. (2021). Computational method on breast cancer survival data using binary classification models. In *Proceedings of the 2nd International Conference on Computational and Bio Engineering*, pp. 107-114. https://doi.org/10.1007/978-981-16-1941-0_12
- [11] Duggento, A., Conti, A., Mauriello, A., Guerrisi, M., Toschi, N. (2021). Deep computational pathology in breast cancer. In *Seminars in Cancer Biology*, 72: 226-237. <https://doi.org/10.1016/j.semcancer.2020.08.006>
- [12] Shawarib, M.Z.A., Latif, A.E.A., Al-Zatmah, B.E.E.D. (2021). Predicting breast cancer diagnosis and survival. *International Journal of Academic Health and Medical Research (IAHMR)*, 5(3): 34-42.
- [13] Nandagopal, V., Geeitha, S., Kumar, K.V., Anbarasi, J. (2019). Feasible analysis of gene expression—a computational based classification for breast cancer. *Measurement*, 140: 120-125. <https://doi.org/10.1016/j.measurement.2019.03.015>
- [14] Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PloS one*, 16(4): e0250370. <https://doi.org/10.1371/journal.pone.0250370>
- [15] Kurama, O. (2021). A new similarity-based classifier with Dombi aggregative operators. *Pattern Recognition Letters*, 151: 229-235. <https://doi.org/10.1016/j.patrec.2021.08.024>
- [16] Maqsood, A., Iqbal, U., Shoukat, I.A., Latif, Z., Kanwal, A. (2021). Fibonacci polynomial based multilayer perceptron neural network for classification of medical data. *AIP Conference Proceedings*, 2355: 040005. <https://doi.org/10.1063/5.0053487>
- [17] Jha, S.K., Pan, Z., Elahi, E., Patel, N. (2019). A comprehensive search for expert classification methods in disease diagnosis and prediction. *Expert Systems*, 36(1): e12343. <https://doi.org/10.1111/exsy.12343>
- [18] Kuhn, K.A., Knoll, A., Mewes, H.W., et al. (2008). *Informatics and medicine. Methods of Information in Medicine*, 47(04): 283-295. <https://doi.org/10.3414/ME9117>
- [19] Gavrishchaka, V.V., Ganguli, S.B. (2003). Volatility forecasting from multiscale and high-dimensional market data. *Neurocomputing*, 55(1-2): 285-305. [https://doi.org/10.1016/S0925-2312\(03\)00381-3](https://doi.org/10.1016/S0925-2312(03)00381-3)
- [20] Aziz, R., Verma, C.K., Srivastava, N. (2018). Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction. *Annals of Data Science*, 5(4): 615-635. <https://doi.org/10.1007/s40745-018-0155-2>
- [21] Soleimanpoor, M. (2020). Discriminating early-and late-stage cancers using multilayer perceptron (Master's thesis, Fen Bilimleri Enstitüsü).
- [22] Banda, S.R.B., Babu, T.R. (2020). A hybrid ensemble feature selection-based learning model for COPD prediction on high-dimensional feature space. In *Data Engineering and Communication Technology*, pp. 663-675. https://doi.org/10.1007/978-981-15-1097-7_55
- [23] Mosayebi, A., Mojaradi, B., Bonyadi Naeini, A., Khodadad Hosseini, S.H. (2020). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PloS one*, 15(10): e0237658. <https://doi.org/10.1371/journal.pone.0237658>
- [24] Lopez-Garcia, G., Jerez, J.M., Franco, L., Veredas, F.J. (2020). Transfer learning with convolutional neural networks for cancer survival prediction using gene-

- expression data. *PloS One*, 15(3): e0230536. <https://doi.org/10.1371/journal.pone.0230536>
- [25] Gupta, P., Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, 171: 593-601. <https://doi.org/10.1016/j.procs.2020.04.064>
- [26] Zorlu, B.Ş.Ç., Kasap, P. (2020). Classification of factors affecting renal failure by machine learning methods. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 36(1): 89-102.
- [27] Ceylan, Z. (2020). Diagnosis of breast cancer using improved machine learning algorithms based on bayesian optimization. *International Journal of Intelligent Systems and Applications in Engineering*, 8(3): 121-130. <https://doi.org/10.18201/ijisae.2020363531>
- [28] Hajiabadi, H., Babaiyan, V., Zabihzadeh, D., Hajiabadi, M. (2020). Combination of loss functions for robust breast cancer prediction. *Computers & Electrical Engineering*, 84: 106624. <https://doi.org/10.1016/j.compeleceng.2020.106624>
- [29] Khaire, U.M., Dhanalakshmi, R. (2020). High-dimensional microarray dataset classification using an improved adam optimizer (iAdam). *Journal of Ambient Intelligence and Humanized Computing*, 11(11): 5187-5204. <https://doi.org/10.1007/s12652-020-01832-3>
- [30] Basavegowda, H.S., Dagneu, G. (2020). Deep learning approach for microarray cancer data classification. *CAAI Transactions on Intelligence Technology*, 5(1): 22-33. <https://doi.org/10.1049/trit.2019.0028>
- [31] Kilicarslan, S., Adem, K., Celik, M. (2020). Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. *Medical Hypotheses*, 137: 109577. <https://doi.org/10.1016/j.mehy.2020.109577>
- [32] Hung, L.C., Sung, S.F., Hu, Y.H. (2020). A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack. *Applied Sciences*, 10(18): 6337. <https://doi.org/10.3390/app10186337>
- [33] Rizal, A., Handzah, V.A.P., Kusuma, P.D. (2022). Heart sounds classification using short-time Fourier transform and gray level difference method. *Ingénierie des Systèmes d'Information*, 27(3): 369-376. <https://doi.org/10.18280/isi.270302>
- [34] Haberman, S.J. (1976). Generalized residuals for log-linear models. In *proceedings of the 9th International Biometrics Conference*, Boston, pp. 104-122.
- [35] Desai, M., Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, 4: 1-11. <https://doi.org/10.1016/j.ceh.2020.11.002>
- [36] Luque, A., Carrasco, A., Martín, A., de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91: 216-231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- [37] Saxena, S., Gyanchandani, M. (2020). Machine learning methods for computer-aided breast cancer diagnosis using histopathology: A narrative review. *Journal of medical imaging and radiation sciences*, 51(1): 182-193. <https://doi.org/10.1016/j.jmir.2019.11.001>
- [38] Guo, B., Gunn, S.R., Damper, R.I., Nelson, J.D. (2008). Customizing kernel functions for SVM-based hyperspectral image classification. *IEEE Transactions on Image Processing*, 17(4): 622-629. <https://doi.org/10.1109/tip.2008.918955>