An optimized K-Nearest Neighbor based breast cancer detection

Tsehay Admassu Assegie
College of Engineering and Technology, Department of Computing Technology, Aksum University, Aksum, Ethiopia
Email: tsehayadmassu2006@gmail.com

***Abstract*— In this research, a grid search is employed to find the optimal hyper-parameter and an optimized K-Nearest Neighbor (KNN) based breast cancer detection model is proposed. The grid search is employed to find the best value of K that could produce better breast cancer detection accuracy. Moreover, this study explored the effect of hyper-parameter tuning on the performance of KNN for breast cancer detection. The findings of this research reveals that hyper-parameter tuning has a significant effect on the performance of the KNN model. The effect of hyper-parameter tuning on the performance of KNN algorithm is experimentally tested using Wisconsin breast cancer dataset collected from kaggle data repository. Finally, we have compared the performance of the KNN with the tuned hyper-parameter and with default hyper-parameter. The result analysis on the performance of the model on breast cancer detection using the testing set reveals that the accuracy of the proposed optimized model is 94.35% and the performance of the KNN with the default hyper-parameter is 90.10%.**

Keywords—breast cancer detection, KNN, optimized KNN, breast cancer, machine learning, hyper-parameter tuning

I. Introduction

Breast cancer is one of the most common types of cancer in the world and the breast cancer causes death [1-31]. To reduce the mortality rate caused by breast cancer, machine learning plays great role in breast cancer identification process. Machine learning algorithms are applied to develop an intelligent system which can identify breast cancer in the early stage as possible in order to reduce the complications and increase survival rate of the patients.

Although machine learning is widely applied in medical research for identification of disease or disease prediction, the performance of the machine learning algorithms needs to be of highest possible level so that the disease can be identified accurately and with higher precision [1-15]. However, without applications of optimization approaches such as the hyper-parameter tuning, the inherent machine learning algorithms cannot predict disease with higher performance [3-6].

Literature review on breast cancer identification with machine learning shows number of works are conducted [1-15] on breast cancer identification problem using machine learning algorithms. But the challenge with the machine learning models is choosing the hyper-parameter that could produce better prediction performance. In this study, hyper-parameter tuning with the help of grid search is employed to develop an optimized K-Nearest Neighbor based model for breast cancer detection with maximum possible performance. Furthermore, this research is focused on investigating the answers to the following research questions:

1) How to optimize KNN algorithm for breast cancer detection?
2) What is the best hyper-parameter or K value that could produce maximum possible performance on breast cancer detection with KNN algorithm?
3) What is the effect of hyper-parameter tuning on the performance of KNN algorithm for breast cancer detection?
4) How to choose the K value for training K-Nearest Neighbor for breast cancer detection?

II. Literature Review

In this section, the literature is surveyed to understand the current state of the art and to identify the problem of breast cancer identification. Numerous researches work [1-15] have been conducted on the identification of breast cancer with machine learning algorithms [1-16]. But the researchers have applied different machine learning algorithms on different breast cancer data repositories and the performance of the proposed model with different machine learning algorithm varies based on the algorithm and the dataset used by different researchers.

In [7], breast cancer detection model is proposed by employing three machine learning algorithms, namely Naïve Bayes, random forest and K-Nearest Neighbor (KNN). The authors have applied these algorithms on the Wisconsin breast cancer data repository. In their study, the authors compared the performance of the proposed model and experimental result shows that the K-Nearest Neighbor (KNN) has better performance than a random forest and the Naïve Bayes algorithm.

In another research on breast cancer identification problem [8], breast cancer prediction model is proposed using decision tree algorithm. The proposed model has an acceptable level of performance for breast cancer detection although, the performance of the model can be improved to get better result for breast cancer detection.

In [9], decision tree-based breast cancer classification model is proposed. the proposed model has acceptable performance for breast cancer identification or breast cancer detection. The experimental analysis on the test set shows that a decision tree algorithm has performed well for breast cancer detection. The performance of the model using accuracy as performance metric is 80.5% for breast cancer detection.

In [10], an artificial neural network (ANN) based breast cancer diagnosis model is proposed. The model has a single hidden layer. The performance of the proposed model is developed using Wisconsin's breast cancer data repository and the model is tested on this dataset.

In [11], K-means clustering is applied to university of California (UCI) breast cancer dataset and a learning model is proposed for breast cancer detection. The performance of the proposed K-means based clustering model is analyzed and result shows that the model has an accuracy of 73.70% on breast cancer detection.

In [12], the performance of decision tree, Naïve Bayes and logistic regression algorithms is compared for breast cancer detection using UCI breast cancer Repository. The comparison on the performance of a decision tree, Naïve Bayes and logistic regression shows that decision tree algorithm has better performance for breast cancer detection.

In [13], neural network-based breast cancer detection model is proposed. The neural network is trained using the (University of California Irvine (UCI) breast cancer data repository. The authors have compared the proposed model with K-Nearest Neighbor and Naïve Bayes. A comparison on the performance of K-Nearest Neighbor, neural network and Naïve Bayes algorithm shows that the neural network has better classification performance than the K-Nearest Neighbor and Naive Bayes algorithm.

In another research on breast cancer identification problem [14], weighted decision tree-based breast cancer identification model is proposed. The authors evaluated the proposed model on test set and result reveals the performance of the model is 94.03%.

In [15] deep learning is applied to Wisconsin's breast cancer dataset and a model for breast cancer is proposed. The proposed model is effective in breast cancer classification although the authors did not mention how hyper-parameters are selected in training phase. The performance of the presented model with deep learning is 90%.

In another research [16], support vector machine (SVM) is employed to automate breast cancer identification. The proposed model is evaluated and result shows the performance of the proposed model is 87.12%. In their study, the authors did not mention how the hyper-parameters are selected for training the support vector machine.

III. Research method

In this section, the method for data collection, best hyper-parameter selection approach employed to find the best K value for the K-Nearest Neighbor and the metrics used in performance evaluation of the proposed model is discussed.

The dataset used in this research is collected from Wisconsin's breast cancer data repository. The dataset consists of 569 observations of malignant and benign cases or observation. The 212 of the observation are malignant or cancerous observations and the 357 observations are benign or non-cancerous observations. Each observation in the dataset have 6 features. The features are shown in figure 1.

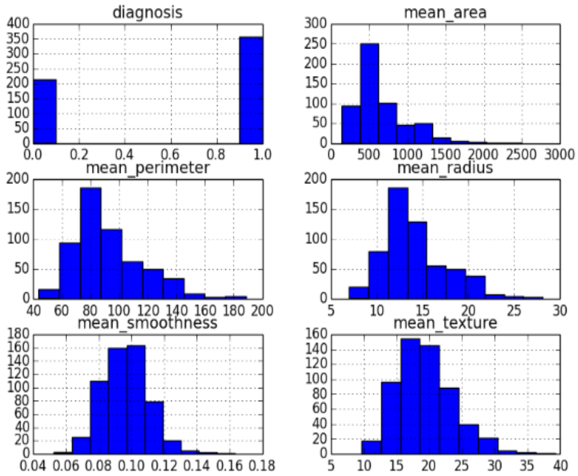

Fig. 1. The features of breast cancer dataset

A. Dataset description

In breast cancer dataset used in this research have six attributes as demonstrated in figure 1. Moreover, the dataset features are illustrated in table 1. The diagnosis feature is used as the class label, indicating the class that a particular observation in the dataset belongs to and other features are used along with the diagnosis feature when the model is trained on the training set. In this research, we have used 70% of the dataset for training and 30% of the dataset for testing the proposed model.

Table 1. Features of breast cancer dataset

Feature	**Description**
Diagnosis	Cass label 0 for benign and 1 for malignant)
Mean perimeter	The mean perimeter of the tumor
Mean smoothness	The variations in radius length
Mean radius	D8stance from center point on the perimeter
Mean area	The mean area of the tumor
Mean texture	Consistence of the surface of tumor

IV. Results and discussions

A grid search approach is used search for best K value. The grid search helps us top lot the accuracy of the KNN against misclassification error. The performance of the proposed model is evaluated using accuracy as a measure of performance on breast cancer test set. The performance of the proposed model is evaluated for different value of K and result is demonstrated in figure 2. As demonstrated in figure 5, the miss-classification error is highest at K value 8 and 39. The miss-classification error is lowest when k value is 9 which shows that the highest possible accuracy achieved by the proposed model 94.35% with miss-classification error value 0.065 as demonstrated in Figure 2.

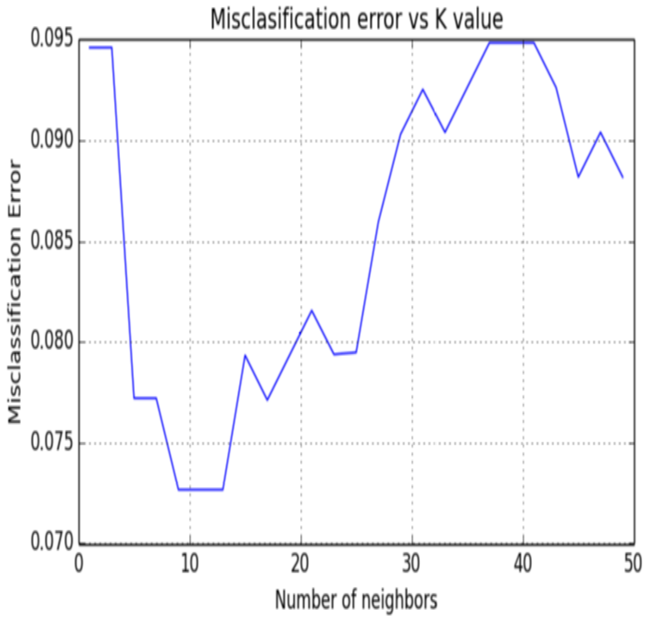

Fig.2 Accuracy vs K values

CONCLUSION

In this research, an optimized KNN model is proposed for breast cancer prediction using a grid search approach for searching the best hyper-parameter. A comparison between default hyper-parameter and tuned hyper-parameter performance is carried out and the result shows that the performance significantly improves when best hyper-parameter or K value is used for training the KNN. The performance of the KNN with default parameters is 90.10%. Better breast cancer detection accuracy is achieved by the KNN using best hyper-parameter or K value is chosen using a grid search approach when the algorithm is trained and the highest performance achieved using hyper-parameter tuning is 94.35%.

REFERENCES

[1] P. Sahity Anaray Anan, Identification of Breast Cancer Using The Decision Tree Algorithm, proceedings of international conference on systems computation, automation and networking, IEEE, 2019.

[2] R. Arulmozhiyal and K. Baskaran, "Implementation of a Fuzzy PI Controller for Speed Control of Induction Motors Using FPGA," Journal of Power Electronics, vol. 10, pp. 65-71, 2010.

[3] Shahrbanoo Goli, Hossein Mahjub, Javad Faradma, Hoda Mashayekhi, Ali-Reza Soltanian, Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression, Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2016.,

[4] James Bergstra, Yoshua Bengio, Random Search for Hyper-Parameter Optimization, Journal of Machine Learning Research, 2012.

[5] Tsehay Admassu Assegie, Pramod Sekharan Nair, The Performance Of Different Machine Learning Models On Diabetes Prediction, International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020.

[6] Ashutosh Kumar Dubey, Umesh Gupta, Sonal Jain, Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data, international journal on advanced science engineering and information technology, 2018.

[7] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury, Breast Cancer Detection Using Machine Learning Algorithms, International Conference on Computational Techniques, Electronics and Mechanical System, 2018.

[8] Mr.P .Sathy Anaranan, proceedings of international conference on systems computation automation and networking, IEEE, 2019.

[9] Shofwatu Uyun, Lina Choridah, Feature Selection Mammogram based on Breast Cancer Mining, International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 1, February 2018.

[10] Mohamed NEMISSI, Halima SALAH, Hamid SERIDI, Breast cancer diagnosis using an enhanced Extreme Learning Machine based-Neural Network, IEEE, 2019.

[11] Tanaya Padhi, Praveen Kumar, Breast Cancer Analysis Using WEKA, IEEE, 2019.

[12] Dr. S. N. Singh, Shivani Thakral, Using Data Mining Tools for Breast Cancer Prediction and Analysis, International Conference on Computing Communication and Automation, IEEE, 2018.

[13] Amandeep Kaur, Prabhjeet Kaur, Breast Cancer Detection and Classification using Analysis and Gene-Back Proportional Neural Network Algorithm, International Journal of Innovative Technology and Exploring Engineering, 2019.

[14] Tsehay Admassu Assegie, Sushma S J., Dr. Prasanna Kumar S C, Weighted Decision Tree Model for Breast Cancer Detection, Technology reports of Kansai university, Volume 62, Issue 03, January, 2020.

[15] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Brwast Cancer Diagnosis by using k-Nearest Neighbor, International Journal of Computer Applications (0975 - 8887) Volume 62 - No. 2013.

[16] Ahmed M, Abdel-Zaher, Ayman M. Eldeib, Breast cancer classification using deep belief networks, expert systems with application, Elsevier, 2016.

[17] Tsehay Admassu Assegie, Sushma S. J, A Support Vector Machine and Decision Tree Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020.

[18] Adel S. Assiri, Saima Nazir, Sergio A. Velastin, Breast Tumor Classification Using an Ensemble Machine Learning Method, journal of imaging, 2020.

[19] M.Chithra Devi, S.Audithan, Breast cancer using ensemble classification and extended weighted voting method International Journal of Advanced Research in Computer Science, 2017.

[20] R.Chtihrakkannan, P.Kavitha, T.Mangayarkarasi, R.Karthikeyan, Breast Cancer Detection using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-11, September 2019.

[21] B.M.Gayathri, ,C.P.Sumathi, T.Santhanam, breast cancer diagnosis using machine learning algorithms: a survey, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.

[22] Y.Ireaneus Anna Rejani, early detection of breast cancer using svm classifier technique, international Journal on Computer Science and Engineering Vol.1(3), 2009.

[23] Basa Varaj Highermath, Sc Prasann Kumar, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), ISSN(P): 2249-6831; ISSN(E): 2249-7943, Vol. 5, Issue 1, Feb 2015.

[24] P. Hamsagayathri, P. Sampath, Performance Analysis Of Breast Cancer Classification Using Decision Tree Classifiers, International Journal of Current Pharmaceutical Research, 2017.

[25] Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta, Diagnosis of Breast Cancer using Decision Tree Models and SVM, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 03 | Mar-2018.

[26] E. Venkatesan, T. Velmuruga, Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, Indian Journal of Science and Technology, Vol 8(29), IPL0625, November 2015.

[27] Alaa. M. Elsayad, H. A. Elsalamony, Diagnosis of Breast Cancer using Decision Tree Models and SVM, International Journal of Computer Applications (0975 – 8887), Volume 83 – No 5, December 2013.

[28] Mevlut Ture a, Fusun Tokatli, Imran Kurt, Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients, Expert Systems with Applications 36 (2009) 2017.

[29] Yuehui Chen, Ajith Abraham, Bo Yang, Feature Selection and

Classification using Flexible Neural Tree, Elsevier science, 2006.

[30] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood, Random Forests and Decision Trees, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[31] Mr. Chintan Shah, Dr. Anjali G. Jivani, Comparison of Data Mining Classification Algorithm for Breast Cancer Prediction, research gate, 2015.