A Hybrid Model for Prediction of Malignancy in Breast Cancer Dataset

Sathiya Devi S [1], *Parthasarathy G* [2]

[1]Department of Information Technology, Anna University BIT Campus, Tiruchirappalli, India

[2]Department of Computer Science and Engineering, SRM TRP Engineering College, Tiruchirappalli, India

[1]sathyadevi.2008@gmail.com

[2]parthasarathee.g@gmail.com

Abstract*— Cancer is the deadliest disease in human beings. The detection of cancer has not become cumbersome with available scientific methods. However, early stage detection of cancer cells is still tedious and inaccurate. The latest advancements in computer and medicinal sciences have paved various new methods in the detection of cancer. One of the latest researches, Machine Learning has been considered for study by various researchers. Supervised machine learning algorithms, works similar to a human, by studying the historical data, which has the ability to read complex datasets and provide many efficient results compared to conventional methods. The aim of this study involves the application of supervised learning algorithms to assist in the detection of cancer cells in the human body. The work considers a voluminous data set comprising of numerical values. Further, this research compares the hybrid model with KNN (K- Nearest Neighbour), MLP (Multi-Layer Perceptron), DT (decision tree), GNB (Gaussian naïve Bayesian) and SVM(Support Vector Machine) classifiers for accuracy. Our study concludes Hybrid Model provides better results while comparing to other algorithms in cancer detection.

***Keywords*—** *prediction, supervised learning, breast cancer, ensemble, classification*

I. INTRODUCTION

Cancer has become a huge burden on various nations in the world. Cancer is a disease that can justifiably receive our full attention in efforts to eradicate it [1]. Cancer can affect all living cells in the body, at all ages, and in both genders. One can readily understand that the cancer cells will soon demand essentially all the nutrition available to the body or to an essential part of the body. As a result of normal tissues gradually suffer nutritive death.

The global increase in the cancer burden and its disproportionate impact on economically developing countries is being propelled by both demographic changes in the populations at the risk and by temporal and geographic shifts in the distribution of major risk factors. The three most important factors that contribute to these trends are viz., Growth and aging of populations, the enrichment of modifiable risk factors and the slower decline in cancer-related to infectious etiology in low resource countries than high resource countries [2, 3].

Early detection and curing of cancer are undergoing research by many researchers [4]. The advancement in biological and computer science, many research studies on computer-aided techniques are under implementation [5]. In recent studies, machine learning (ML) approaches provide a significant method for diagnosis and treatment of cancer [6].

Machine learning is a technique that learns the historical data and predicts the values. This technique is available for categorical data and continuous data. The learning activity is classified into two main types (i) Supervised (ii) Unsupervised. In Supervised learning, the target label is known and not in Unsupervised learning [7].

The aim of this paper is to detect the malignancy by applying a numerical dataset as input to the various classification models. These models are evaluated to measure the accuracy of the prediction. In the evaluated model squad, high-performance models are combined as a hybrid model and evaluated.

The rest of the paper is organized as follows: Section 2 describes the literature review for the Hybrid Recommendation system. The proposed approach for improving the accuracy of the recommender system is presented in section 3. Section 4 discusses the experiment and the results of the proposed method. The conclusion has arrived in section 5.

II. LITERATURE REVIEW

Cancer is the leading cause of death in developed countries and is rising in alarming rates in developing countries. The global burden of cancer continues to increase largely because of the increase in longevity and growth of the world population alongside an increasing adoption of cancer-causing behaviours, particularly smoking, in economically developing countries. The GLOBOCAN 2018 estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer, with a focus on geographic variability across 20 world regions. There will be an estimated 18.1 million new cancer cases and 9.6 million cancer deaths in 2018[8].

In India cancer is the major killing disease. The total number of new cases in 2018 is 1.15 million cases including both sexes and all ages. Males of all ages are 0.57 million cases and Females of all ages are .58 million cases [9].

Related literature for classification approach is as follows.

Way et al. developed a machine-learning approach using PanCanAtlas data to detect Ras activation in cancer. Integrating mutation, copy number, and expression data, the authors show that their method detects Ras-activating variants in tumours and sensitivity to MEK inhibitors in cell lines [10].

An SVM classifier based model is proposed to classify all types of cancer both common and rare. In this integrated model, features are extracted by a natural language processing from death certificates and features exploited using machine learning classifiers on death certificates. It also combines machine learning and rule-based methods to classify the documents [11].

Ritcher et al. Proposed a prediction model using machine learning techniques such as Decision Trees, Neural Networks, and Support Vector Machines. This study focuses on utilizing structured clinical information and this data is widely collected and has the greatest value for efficient modelling of cancer risk and recurrence [12].

An ensemble system has developed a set of individually trained classifiers whose decisions are combined typically with majority voting, weighted voting or other relatively simple techniques such as stacking or Naïve Bayes combination. Researches show that generally ensemble classifier outperforms the performance of the best member classifier in the squad [13].

Hasan et al. proposed a machine learning based classification of cancer cells based on cell gestures [14]. The identification scheme was validated with untrained data and three different classifiers were used to construct the system and compare the performances. In this work, the Naïve Bayesian classifier identified the cancer cells with the highest accuracy.

Motivated by the above literature, supervised machine learning approaches play a vital role to detect the cancer cells. This paper proposes a hybrid machine learning models to detect breast cancer cells. This early detection will be useful to cure cancer. The proposed method is explained in the next section.

III. PROPOSED SYSTEM

This paper fulfils the gap by introducing numerical datasets as input data to detect cancer using various supervised learning algorithms. In order to implement this proposed model, we used python platform where large datasets handled easily. The architecture of the proposed model contains (i) Data Preparation, (ii) Feature Engineering, (iii) Model Creation, (iv) Prediction, (v) Evaluation (Fig. 1)

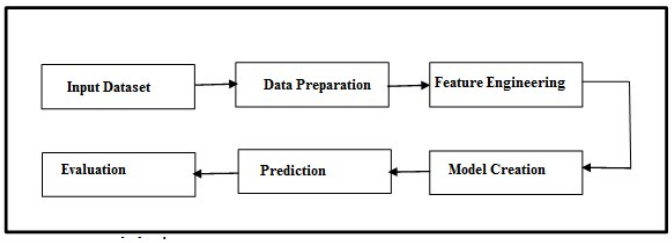

Fig. 1 Proposed System

A. Data Preparation

Preprocessing is used to remove inconsistency and noise in the data since it affects the accuracy of the result. These barriers in the dataset should be handled by various preprocessing techniques like data cleaning, data integration, data reduction and data transformations [15]. Pre-processing is a process of transform raw data into an understandable format. The data set is checked for the missing value, null value, and duplicate value. All the categorical values are converted into a numerical value. This process can be achieved through label encoder, where encoder converts categorical or text into numerical values. The 10 fold cross-validation is used to split the train and test set

B. Feature Engineering

Feature engineering is the process of creating features using domain knowledge of the data that improve machine learning algorithms to work more efficiently [16]. The continuous features become identical in terms of the range, after a scaling process. The common way of scaling is the normalization or min-max normalization where all values in a fixed range between 0 and 1, which is done by the following equation.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

C. Model Creation

A model has developed by using the various classification algorithms like k Nearest Neighbour, Multi-Layer Perceptron, Support Vector Machine, Decision Tree and Gaussian naïve Bayesian. These models are trained by using the training dataset. In this work, hybrid model is created by combing the top two classifiers based upon the precision results.

D. Prediction

The trained dataset is used to create the different classification models. Prediction is the process of measuring the class label using other features in the dataset. In this work, the malignancy presence is predicted using other features.

E. Evaluation

The proposed model is validated using the performances measure like precision, recall and F1 measure. These measures are calculated by using the confusion matrix. The sample confusion matrix in the experiment is described in the Table 1.

TABLE I CONFUSION MATRIX

n=35	**Predicted : Yes**	**Predicted: No**	**Total**
Actual: Yes	TP=9	FN=2	11
Actual: No	FP=7	TN=17	24
Total	16	19	35

The formula to compute precision is mentioned in the following equation. Precision or Confidence denotes the proportion of Predicted Positive cases that are correctly Real Positives.

$$Precision = \frac{tp}{tp+fp} \quad (2)$$

Recall or Sensitivity is the proportion of Real Positive cases that are correctly Predicted Positive. The formula used to find the recall is mentioned below.

$$Recall = \frac{tp}{tp+fn} \quad (3)$$

F1 is a function of Precision and Recall. The formula used to find the recall is mentioned below.

$$f1 = 2\ \frac{precision*recall}{precision+recall} \quad (4)$$

IV. EXPERIMENTS AND RESULT

The input dataset used for the experiment is the numerical dataset UCI Machine Learning Repository. The dataset corresponding to breast cancer Coimbra dataset is considered for the experiment. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis[17].

The data set is split into 2 sets which are training and testing set .70% of data is split for training and the remaining 30 % for testing using 10 fold cross-validation. Based on the comparison of test and train data, a training model will be created. The trained classification model is used to predict the presence of breast cancer.

There are two stages of evaluation in this proposed model. The models are created using a training dataset and values are predicted. The performance measures like confusion matrix, precision, recall, and F1 measure are found. Table 2 shows the performance measures of different models before applying the feature engineering technique. Precision is high in the hybrid model which is created by combining the decision tree and Gaussian Bayesian Models. In the independent models, the decision tree model outperforms while comparing with other classification models shown in Fig. 2.

TABLE II. PERFORMANCE MEASURES OF MODEL BEFORE FEATURE ENGINEERING

Model	Precision	Recall	F1
KNN	0.57	0.57	0.56
MLP	0.33	0.34	0.33
SVM	0.59	0.56	0.53
DT	0.75	0.75	0.75
GB	0.74	0.72	0.71
Hybrid Model	0.77	0.72	0.7

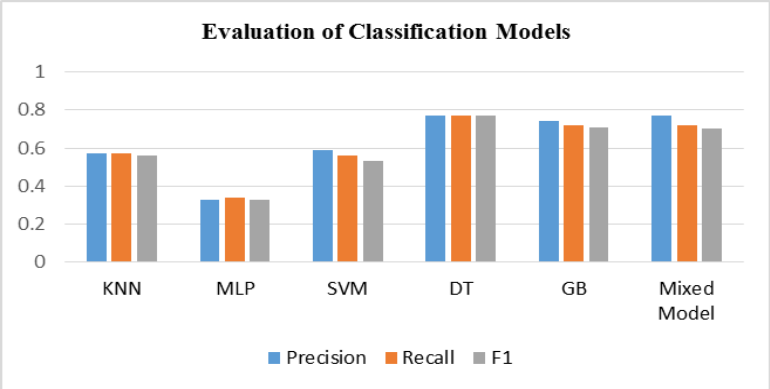

Fig. 2. Comparison of Models before Feature Engineering

Table 3 shows the performance measures of different models after applying the feature engineering technique. Precision is high in the Hybrid model which is created by combining the decision tree and Gaussian naïve Bayesian while comparing with other models. The performance comparison of Models shown in Fig.3.

TABLE III. PERFORMANCE MEASURES OF MODEL AFTER FEATURE ENGINEERING

Model	Precision	Recall	F1
KNN	0.63	0.61	0.63
MLP	0.77	0.76	0.76
SVM	0.56	0.56	0.55
DT	0.73	0.76	0.73
GB	0.71	0.74	0.72
Hybrid Model	0.77	0.8	0.78

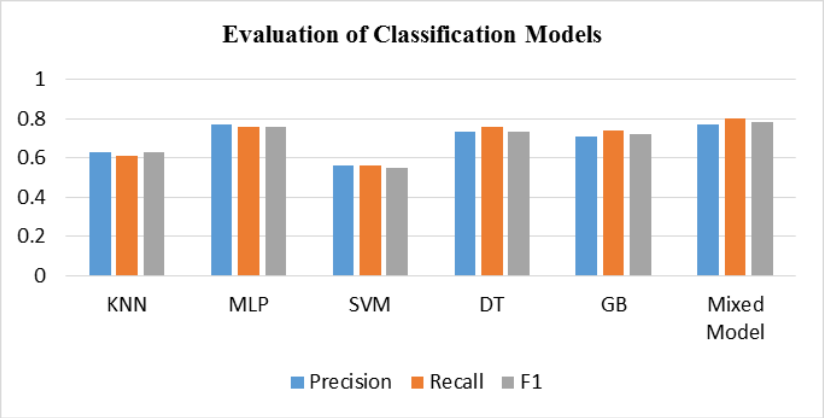

Fig. 3. Comparison Models after Feature Engineering

V. CONCLUSIONS

Machine learning plays a major role in the medicinal and scientific field. This paper is to detect the cancer cell using supervised learning algorithms. The hybrid model is also developed using the basic classifier algorithms, which shows higher accuracy. This system is implemented in the breast cancer dataset using five supervised machine learning algorithms. The various performance measures are recorded to evaluate the system. The results show that Hybrid which combines Decision Tree and Gaussian Naïve Bayesian produced better results. The feature engineering technique is applied in the dataset and models are created using the same classifiers. The same evaluation steps are followed to measure the models and recorded. The results show that the hybrid model gives better results. This work concludes that the hybrid model gives a high impact on accuracy improvement. Feature Engineering Technique is given high accuracy while comparing with the base models.

REFERENCES

[1] Chanda S, Nagani K, " In vitro and in vivo Methods for Anticancer Activity Evaluation and Some Indian Medicinal Plants Possessing Anticancer Properties: An Overview", *J Pharmacog Phytochem*, pp 140-152,2013

[2] Thun MJ, Henley SJ, Burns D, Jemal A, Shanks TG, Calle EE, " Lung cancer death rates in lifelong nonsmokers". *J Natl Cancer Inst*, pp 691-699,2006

[3] Nair MK, Varghese C, Swaminathan R, "Cancer: Current Scenario, Intervention Strategies and Projections For NCMH Background Papers. Burden of Disease in India",pp 219-225,2015

[4] D. M. Joshi, N. K. Rana, and V. M. Misra, "Classification of brain cancer using artificial neural network," *ICECT 2010 - Proc. 2010 2nd Int. Conf. Electron. Comput. Technol.*, pp. 112–116, 2010.

[5] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images*," Procedia Comput. Sci*., pp. 107–114, 2018,vol.125.

[6] Tapak L, shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J, " Prediction of survival and metastasis in breast cancer patients using machine learning classifiers",*Clinical Epidemiology and Global Health* (2018).

[7] C.C Aggarwal," Recommender Systems: The Textbook", Springer International Publishing Switzerland 2016.

[8] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram,Rebecca L. Siegel, Lindsey A. Torre, Ahmedin Jemal, "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA CANCER J CLIN* 2018.

[9] India Fact Sheets,[Online].Available:http://gco.iarc.fr/today/data/factsheets/populations/356-india-fact-sheets.pdf

[10] Way et al., 2018, Cell Reports 23, 172–180 April 3, 2018 ,[Online].Available: https://doi.org/10.1016/j.celrep.2018.03.046

[11] Koopman, B., Artificial Intelligence In Medicine (2018), [Online].Available:https://doi.org/10.1016/j.artmed.2018.04.011

[12] Richter, A.N., Artificial Intelligence In Medicine (2018), [Online],Available:https://doi.org/10.1016/j.artmed.2018.06.002

[13] Tarek S et al. "Gene expression based cancer classification". *J Egyptian Informatics* (2016), http://dx.doi.org/10.1016/j.eij.2016.12.001

[14] Mohammad R. Hasan , Naeemul Hassan , Rayan Khan , Young-Tae Kim , Samir M. Iqbal , Classification of Cancer Cells using Computational Analysis of Dynamic Morphology, Computer Methods and Programs in Biomedicine (2017), doi: 10.1016/j.cmpb.2017.12.003

[15] JiaweiHan MichelineKamber ,JianPei, "Data Preprocessing-Data Mining (Third Edition) The Morgan Kaufmann Series in Data Management Systems", pp 83-124,2012.

[16] Tara Rawat and Vineeta Khemchandani, "Feature Engineering (FE) Tools and Techniques for Better Classification Performance",*International Journal of Innovations in Engineering and Technology (IJIET)*,http://dx.doi.org/10.21172/ijiet.82.024,2017.

[17] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, Accessed on 03.05.2019