PREDICTIVE DATAMINING AND KERNEL PERCEPTION SUPPORT VECTOR MACHINE (KPSVM) ALGORITHM IN MEDICAL DATA

Deepthy Babu.N[1]
Research Scholar
PG and Research Department computer science
Nehru Arts & Science College,
Coimbatore, Tamil Nadu-641105

Dr.N.Kavitha[2]
Head, Associate Professor[2]
PG and Research Department computer science
Nehru Arts & Science College,
Coimbatore, Tamil Nadu-641105

Abstract: - Radiology is a vast subject and requires more knowledge and understanding for exact diagnosis of tumor in medical science. In this work, a meningioma segment and detection approach is designed using sequence dataset as input data for defining the cancer point. This expriments is a difficult to the large diversification in the existence of tumor tissues related to various inmate and most of the cases similarity within the normal tissues makes the task difficult. The main impartial is to categorize the cancer into the presence meningioma or a healthy breast. In this proposed, data mining and statistical learning techniques were applied to cancer datasets for survival analysis. The breast cancer dataset from kaggle machine learning database system were used for prediction and comparative study of the data mining and statistical learning techniques. The results of the classifiers or models were mixed; existing methods Support Vector Machines (SVM), K-means, Logistic regression, neural network and proposed method KPSVM based on accuracy. However, the KPSVM showed higher classification accuracy (91.40%) over neural network, support vector machine, Kmeans and logistic regression models in terms of accuracy.

Keywords:- K-means, kernel perception, breast cancer, machine learning algorithm.

I. INTRODUCTION

Over the most recent two decades, we have been acquainted with another field of study or ideas that have directly affected the combination of software engineering, measurements, science and drug. Biomedical is the term currently connected to using medication and science look into discoveries because of clinical inquiries for enhanced basic leadership. the consequences of a far reaching similar investigation of information mining, measurable and machine learning calculations, including Support Vector Machines (SVM), Artificial Neural Networks (ANN), neural system and K-implies calculations were analyzed. The fundamental focal point of this exploration was to contemplate the viable grouping learning procedures for forecast of bosom malignant growth survivability. There are two fundamental viewpoints in forecast of malignant growth survivability: precision and effectiveness. It is one among the emerging technologies, with its branches of application wide spread into several domains of business. playing out, the extent of the preparation dataset and how well the dataset is considered. The reason for this investigation is to investigate the abilities of factual learning and information mining procedures utilized in organic datasets. Information mining and factual strategies were utilized to investigate a breast cancer growth dataset and the precision and productivity of every system were analyzed. Python and Pandas programming were utilized to investigate the bosom malignancy dataset. Python is open source factual examination programming, and Pandas is open source machine learning application programming that can be utilized to standardize and investigate datasets. Pandas give usage of learning calculations that we can undoubtedly apply to any dataset.

So, there requires an assistant tool which helps in detecting the presence of tumor in the Slices image of breast and stage accurately. Thus detection of tumor in brain plays a curial and tough job in the range of medical picture processing. The separation of damaged or infected part from the cancer along with its shape, size and boundary is known as identification of meningioma.

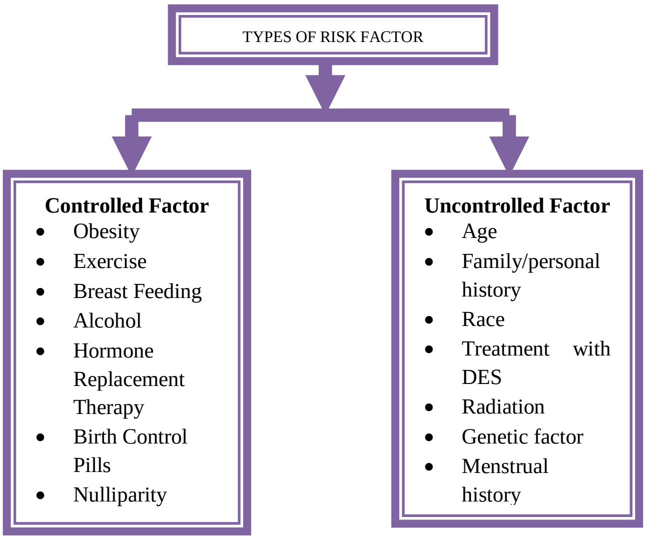

Figure 1 Breast Cancer Risk Factors

The subsist of the paper is described as. The review in literature is done in Section II. The theoretical design and the phases of the scheme are described in Section III and the experimental results of the computerized system are documented in Section IV. Finally, future work and conclusions are written Section V.

II. LITERATURE REVIEW

Breast disease is an uncontrolled development that happens in bosom tissues and the most normal sort is ductal carcinoma, which starts in the coating of the drain pipes. Another kind of bosom disease is lobular carcinoma, which starts in the lobules (drain organs) of the bosom [7] [8]. the ailment happens for the most part in ladies, yet men can get it, as well (Abeloff et al., 2008). [9] The less normal bosom malignancy types are provocative bosom malignant growth, Triple-negative bosom malignant growth, Paget infection of the areola, Phyllodes tumor and Angiosarcoma (Santen and Mansel, 2005). [10] It is assessed that in the USA in 2014 roughly 232,670 instances of female bosom malignant growth will be analyzed, and roughly 40,000 ladies will kick the bucket from the malady (Caplan and Delay, 2014).

Data about past works done by different analysts and similar examination of arrangement and bunching calculations is completed in this segment. The execution measurements of various informational index for medicinal and some other related applications are talked about. The primary objective of this examination work is making conceivable a determination of calculations and dataset classification, for appropriate medicinal applications in future. These days, a lot of medicinal information, putting away patients' medicinal history, is gathered amid social insurance. The examination of these medicinal information gathered, is a testing undertaking for medicinal services frameworks, since huge measure of fascinating learning can be naturally mined to viably bolster both doctors and human services associations. There are quantities of research works completed by various individuals to locate the medicinal information. By and by, there are a number of systems utilized and connected to dissect restorative sicknesses. Here, such systems are represented.

Several methods have been implemented to detect and distinguish the cancer cells for diagnosis. Wang *et.al*(2016) proposed the machine learning algorithms to improve the ability of single hidden layer node and to choose the best input weights. ELM algorithm determines the weight and hidden layer nodes and mainly implemented to reduce the learning speed. SVM is used to simplify the ELM network and to improve the generalization of ELM performance. The hidden layer nodes weights are determined by SVM and then the SVM-ELM model are optimized by particle swarm optimization [1].

Zheng *et.al*(2014) introduced a hybrid of K-means and support vector machine algorithms based on feature extraction for diagnosis of breast cancer. The K-means algorithm is used to determine the hidden patterns of the cancer cells on benign and malignant tumors and these patterns are trained by SVM model. This integration model reduces the computation time with better classification accuracy [2].

Rana *et.al* (2015) describes the survey of machine learning techniques for predicting breast cancer cells. These machine learning techniques are stimulated by MATLAB using UCI machine learning depository. From the experimental results, it is clear that SVM techniques using Gaussian kernel can predict the cancer cells in terms of recurrence/non-recurrence breast cancer [3].

Peña Ayala *et.al*(2014) explained a survey of data mining techniques in the education system. The main aim is to improve the Educational Data Mining (EDM) development by analyzing the review based on the data mining outcomes [4],[7]. Utomo *et.al* (2015) implemented an ANN with extreme learning techniques for diagnosing breast cancer. The ANN delivers a good result but it lacks with long training period due to intricate architecture. It can be solved by combining with ELM with fast learning speed. The integration of ELM and ANN has better generalization performance than BP-ANN for diagnosing breast cancer [5].

Rana *et.al* (2015) describes the survey of machine learning techniques for predicting breast cancer cells. These machine learning techniques are stimulated by MATLAB using UCI machine learning depository. From the experimental results, it is clear that SVM techniques using Gaussian kernel can predict the cancer cells in terms of recurrence/non-recurrence breast cancer [6],[8].

Arrangement systems utilized for the expectation of malignancy stages were additionally examined [9]. Malignant growth level expectation was made to decide the treatment procedure. Research approaches completed in this work are talked about in whatever remains of the parts [7].

III. SYSTEM METHODOLOGY

Many researchers have developed different machine learning techniques to diagnose the disease. Different models have been developed using different algorithms for prediction of the disease. But no precise model has been developed in diagnosing the disease accurately. Hence, in this proposed research, an integrated framework model developed.

Wisconsin Diagnostic Breast Cancer Dataset

The data consists of 569 instances (records) of nuclear features of fine needle aspirates taken from patient's breast and 32 attributes (ID, diagnosis, 30 real-valued input features). Each instance has ten attributes and the class attribute. The basic block diagram for tumor detection and classification consist of a set input data called as dataset, these data are fed to the pre-processing block where the data are smoothened and noise will be removed by the use of different algorithms [8].

Next the dataset undergoes into the process segmentation where the precise boundary is obtained depending on the area of interest with human interaction or by automatic segmentation without human interaction. The segmented region's pixels are extracted by feature extraction algorithms to compare the extracted pixel with the victim case for detecting the brain region is affected or not to classify into severe or benign [9].

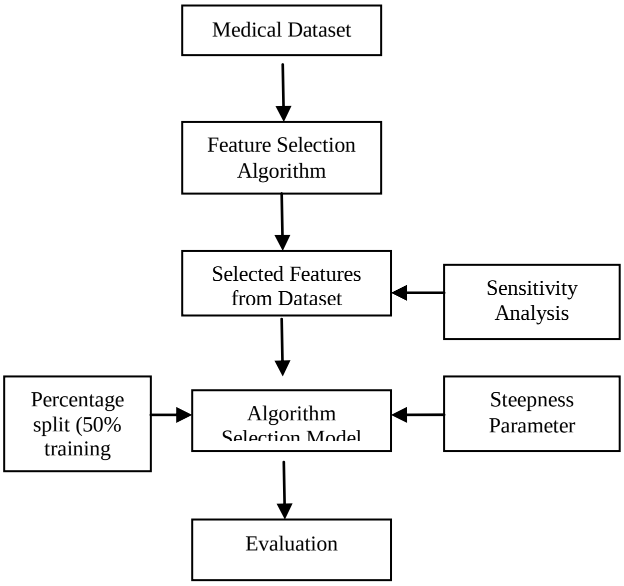
Medical Dataset
Feature Selection Algorithm
Selected Features from Dataset
Sensitivity Analysis
Percentage split (50% training
Algorithm Selection Model
Steepness Parameter
Evaluation

Figure 2 Cancer Prediction Model

Kernel perception Support Vector Machine (KPSVM)

Kernel perception Support Vector Machine (KPSVM) is a streamlining technique in light of an immediate relationship to Darwinian common determination in natural proliferation. In this rule, hereditary calculations are encoding a parallel hunt through idea space, with each procedure endeavoring coarse-grain slope climbing (Goldberg 1988) [12].

Various methodologies have been used in the use of SVM to prediction issues (Davis 1985, Goldberg and Lingle 1985, Starkweather et al 1992)[14]. Instigated varieties and recombination of these ideas are tried against an assessment capacity to see which one will make due to next generation. To reach a minimum of dissimilarity function there are two conditions. These are given in Equation 1 and Equation 2.

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} x_j}{\sum_{j=1}^{n} u_{ij}{}^{m}} \qquad (1)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \qquad (2)$$

Detailed algorithm of KPSVM proposed by Bezdek in 1978.

$$\frac{\partial C}{\partial w^{l}_{jk}} = a^{l-1}_{k} \delta^{l}_{j} \text{ and} \qquad (3)$$

The utilization of hereditary calculations requires the accompanying segments:

- A method of encoding answers for the issue.
- An assessment task that profits a rating for every arrangement.
- A method of instating the number of inhabitants in arrangements.
- Operators that might be connected to guardians when they imitate to change their hereditary structure, for example, hybrid, transformation and other space particular administrators.

Pseudocode for KPSVM performance estimation procedure

```
INPUT: Wisconsin Diagnostic Breast Cancer Dataset
OUTPUT: Classification result
1.dataset with each attributes
2.Find KPSVM parametric value for matrix coordinate
3.Calculate weight least cost
4.Select  Attribute $_{average}$
5. Calculate the value $l_i(t)$.
6.Updatation process depends on node variation
7.find accuracy and cancer type
Return
Calculate weight least cost step 3
```

Figure 3 Pseudocodes for Kpsvm WDBC Prediction Model

In the back propagation classification technique, have been proposed in which supervised classifier works as trained classifier which will compare,

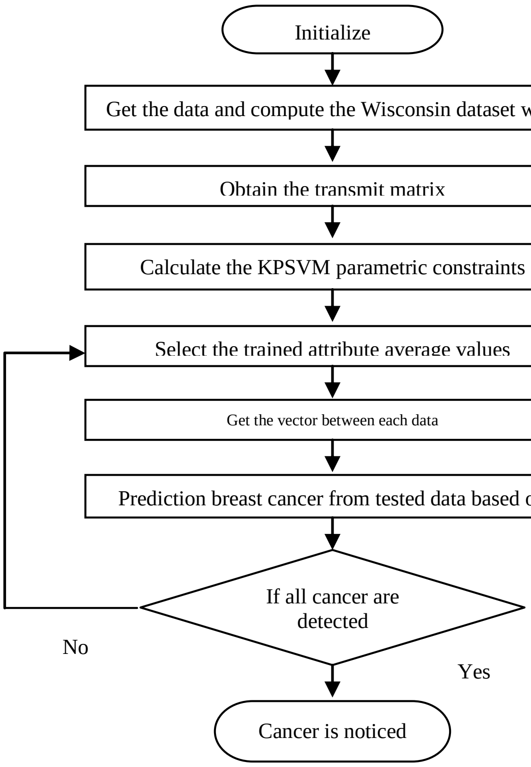
Initialize
Get the data and compute the Wisconsin dataset with
Obtain the transmit matrix
Calculate the KPSVM parametric constraints
Select the trained attribute average values
Get the vector between each data
Prediction breast cancer from tested data based on
If all cancer are detected
No
Yes
Cancer is noticed

Figure 4 Proposed KPSVM model flow chart

The input data with the stored knowledge data base and the unsupervised classifier stores the important features of the respective dataset for comparing with upcoming input data feature. Here we have supervised classifier trained with victim dataset features to compare with feature extracted from the segmented MR image detailed in equation 3.

IV. RESULT AND DISCUSSION

To analyze the performance of the proposed model, accuracy, precision and recall were used for evaluating classification results and mean squared error (MSE) and R2 score were used for evaluating regression results. The Kernel Perception SVM Algorithm was chosen to solve this problem. In figure 5 display confusion values of kpsvm algorithm.

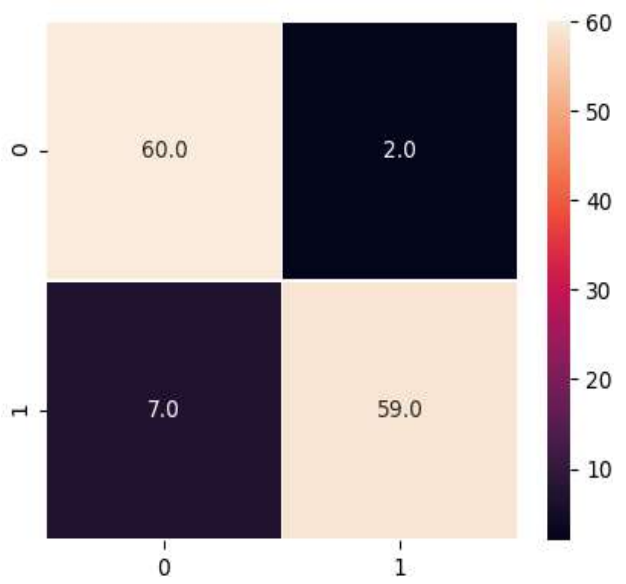

Figure 5 normalized KPSVM confusion matrix predicted values

To obtain the discrimination power and the Support Vector Machine algorithm feature set of Wisconsin Diagnostic Breast Cancer Dataset users are noticed with respect to the generated values as shown in the figures 5 and 6. Total predicted accuracy is 91.40 % so the conclusion was that the model has performed quite well on this independent dataset.

Table 1 Existing Algorithm models compression results

Algorithm	Accuracy /Efficiency	Precision	Sensitivity
Support Vector Machine	**88.88**	**90**	**91**
Neural network	**87.87**	**90**	**87**
Logistic regression	**86.86**	**88**	**88**
K-means	**90.90**	**91**	**96**

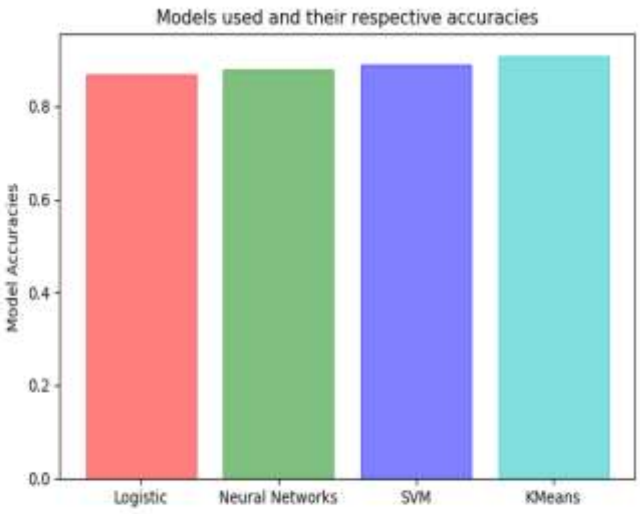

Figure 6 Existing Algorithm Accuracy compression results

Shown in Figure 6, and table 1, In existing method compression Kmeans algorithm is showed slight improvement when applied to the independent dataset. Comparatively Kmeans algorithm slightly improves and best result for other three algorithms (logistic regression, neural network, and support vector machine) in various parameters (accuracy, precision, sensitivity).

Table 2 Proposed Algorithm models compression with existing result

Algorithm	Accuracy /Efficiency
Support Vector Machine	88.88
Neural network	87.87
Logistic regression	86.86
K-means	90.90
Kernel Perception SVM	91.40

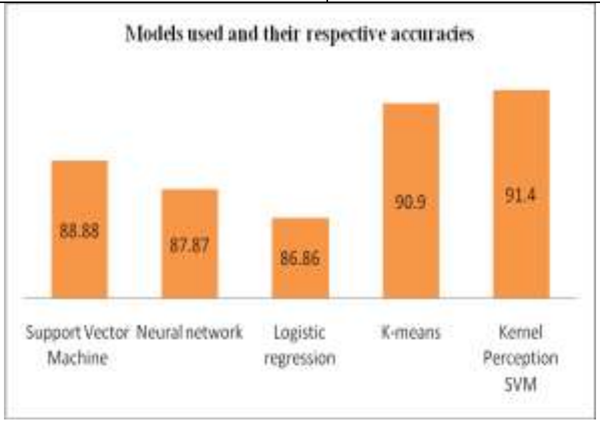

Figure 7 proposed Algorithm Accuracy compression results

Shown in Figure 7, and table 2 Based on the result algorithms are ranked as Kmeans, Support Vector Machine, Neural Network, and Logistic Regression. But Comparatively proposed KPSVM algorithm slightly improves and best result for other existing algorithms (logistic regression, neural network, Kmeans and support vector machine) in various parameters (accuracy, precision, sensitivity). Based on the result algorithms are ranked as KPSVM, Kmeans, Support Vector Machine, Neural Network, and Logistic Regression.

The proposed methodology has maximum accuracy, Average F-score and expected average boundary distance. Hence, it is adaptable to any type of dataset. In future, these concepts are adaptable for each and every real time sequence and verify the test results. At last, the average correlation is achieved as expected. Hence, it is adaptable to any type of brain tumor diseases. In future, these concepts are adaptable for each and every real time image sequence and verify the test results.

V. CONCLUSION

In this proposed work, an attempt has been made to evaluate the performance of proposed algorithms KPSVM, existing algorithms logistic regression, neural network, Kmeans and support vector machine model with feature selection and sensitivity analysis using percentage split on publicly available Wisconsin Diagnostic Breast Cancer Dataset. Here, the regions are partitioned and segmented into various levels WDBC dataset. Accuracy wise, then the best model is Kernel perception support vector machine (91.40%). Results were mixed as to which algorithm is the most accurate model, and it appeared that the performance of each algorithm depends on the size and the cleanliness of the dataset. KPSVM did outperform Kmeans, SVM, logistic regression and neural network algorithms based on accuracy.

Since, the role of Adaptively Regularized KPSVM gives high accuracy of each prediction. From the results it is noticed that Shown in Figure 7, and table 2 Based on the result algorithms are ranked as Kmeans, Support Vector Machine, Neural Network, and Logistic Regression. But Comparatively proposed KPSVM algorithm slightly improves and best result for other existing algorithms (logistic regression, neural network, Kmeans and support vector machine) in various parameters (accuracy, precision, sensitivity). Based on the result algorithms are ranked as KPSVM, Kmeans, Support Vector Machine, Neural Network, and Logistic Regression.

REFERENCES

[1] Wang, M.M. and Ding, S.F., 2016. The SVM-ELM Model Based on Particle Swarm Optimization. In Proceedings of ELM-2015 Volume 2 (pp. 93-105). Springer, Cham.

[2] Zheng, B., Yoon, S.W. and Lam, S.S., 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), pp.1476-1482.

[3] Krishnaiah, V., Narsimha, D.G. and Chandra, D.N.S., 2013. Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies, 4(1), pp.39-45.

[4] Peña-Ayala, A., 2014. Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), pp.1432-1462.

[5] Utomo, C.P., Kardiana, A. and Yuliwulandari, R., 2014. Breast cancer diagnosis using artificial neural networks with extreme learning techniques. International Journal of Advanced Research in Artificial Intelligence, 3(7).

[6] Rana, M., Chandorkar, P., Dsouza, A. and Kazi, N., 2015. Breast cancer diagnosis and recurrence prediction using machine learning techniques. IJRET: International Journal of Research in Engineering and Technology eISSN, pp.2319-1163.

[7] Virnig, B. A., Tuttle, T. M., Shamliyan, T., & Kane, R. L. (2010). Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. Journal of the National Cancer Institute, 102(3), 170-178.

[8] Abeloff, M. D., Wolff, A. C., Weber, B. L., Zaks, T. Z., Sacchini, V., & McCormick, B. (2008). Cancer of the breast. Clinical Oncology. 4th ed. Philadelphia, Pa: Elsevier, 1875-1943.

[9] Santen, R. J., & Mansel, R. (2005). Benign breast disorders. New England Journal of Medicine, 353(3), 275-285.

[10] Caplan, L. (2014). Delay in breast cancer: implications for stage at diagnosis and survival. Frontiers in public health, 2, 87.

[11] J. M. Moguerza and A. Muñoz "Support vector machines with applications" Statistical Science pp. 322-336 2006.

[12] L. Ya-qin, W. Cheng, and Z. Lu, “Decision tree based predictive models for breast cancer survivability on imbalanced data,” pp. 1-4, 2009.

[13] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artif. Intell. Med., vol. 34, pp. 113-127, 2005.

[14] Pradesh, “Analysis of Feature Selection with Classification: Breast Cancer Datasets,” Indian J. Comput. Sci. Eng., vol. 2, no. 5, pp. 756-763, 2011.

[15] El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F. and Al-Suhaim F., “Evaluation of breast cancer tumor classification with unconstrained functional networks classifier,” Computer Systems and Applications, IEEE International Conference, 2006, pp. 281 – 287.

[16] Sudhir D., Ghatol Ashok A., Pande Amol P., “Neural Network aided Breast Cancer Detection and Diagnosis”,7th WSEAS International Conference on Neural Networks, 2006