Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777251

CLASSIFICATION HYBRID ALGORITHM BASED ON SVM AND KNN WITH MODIFICATIONS

K. Saranyadevi
Research scholar, Navarasam Arts and science college for women, Arachalur, Erode dt Tamilnadu India

Dr. P. Rathiga
Assistant professor, PG and research development Erode arts and science college, Erode Tamilnadu India

Abstract
Digital mammography is the method that has proven to be the most trustworthy and productive in terms of early and precise detection of breast cancer. Within the realm of medicine, the diagnosis and categorization of breast cancer are both helped along significantly by image processing. In this article, a system is developed to classify mammography pictures into three distinct categories: benign, malignant, and normal. These categories are discussed more below. Mammogram images go through a preliminary processing step, during which the segmented region's features are extracted. These features are input into a modified version of the SVM and a KNN classifier to be trained. In order to assist in the classification of the mammography pictures, the authors propose a hybrid approach that combines the modified SVM and KNN classifiers. The most recent method, which is an improvement on the SVM algorithm, is one that introduces multi class for the classification of breast cancer. Utilising the KNN method in accordance with the distribution of test images inside a feature space, it does this. In addition to that, the SVM and KNN classifiers' degrees of accuracy are evaluated here. The accuracy of the prognosis produced by the modified SVM and KNN hybrid algorithm is superior to that produced by either the KNN approach or the SVM methodology. Using 10 test photos and 20 taught ones, this approach was evaluated. When it comes to the classification of mammography pictures, our system obtains an overall mean accuracy of 99.3406%. Classification, KNN, MIAS, and Proposed KNN with SVM are some of the keywords here.

I. INTRODUCTION

A tumour that develops in the breast cells is what we refer to as breast cancer. It is the most common type of tumour in women that does not occur on the skin, and it is the second most common cause of sickness in women. Breast cancer survival rates are higher than they were in the past, and the number of deaths attributed to this disease is constantly decreasing. This is partly attributable to the earlier diagnosis of tumours as well as other reasons. On mammograms, breast tumours and masses almost always appear as dense patches. This is because they are. A malignant tumour, on the other hand, will typically have a boundary that is irregular, rough, and blurry. A benign mass, on the other hand, will typically have a round, smooth, and well constrained boundary.

It is critical to detect cancer at an early stage in order to provide a prompt response and improve treatment prospects. Unfortunately, early cancer detection can be challenging due to the absence of symptoms associated with the disease in its early stages. Therefore, cancer continues to be one of the subjects of research in the field of health, and numerous researchers have contributed to this field in the hope of producing evidence that can lead to the development of therapy, prevention, and diagnostics.

Machine learning encompasses two separate approaches to education: supervised and unsupervised. First, specify the classes that will be used to categorise the existing data; next, declare the classes that will be used in the next phase. Support Vector Machines, Decision Trees, Neural Networks, Bayesian Networks, k-Nearest Neighbours, and many more are just some of the many machine learning methods out now.

Classifying data sometimes involves employing a technique called k-nearest neighbour. In the KNN, a new component is assigned a classification based on its distance from an already established component. For the system to function properly, it must have 2956.

depending on the distance specified and the parameter k that determines the number of neighbours used to determine the new element's classification. These two elements need one another to function properly.

The Support Vector Machine (SVM) is a popular machine learning technique that has proven effective in many different settings. Its ability to generalise makes it useful for tasks such as data classification and function estimate. By increasing the distance between the dataset and the hyper plane, the Support Vector Machine can reduce both the lower bound and the upper bound of the generalisation error. Since the optimal number and locations of the fundamental functions may be mechanically obtained by training, the Support Vector Machine provides an extra benefit over conventional model selection. That's not how most model selection procedures work. The success of SVM depends heavily on the kernel.

The method used to categorise digital mammography pictures is investigated here. Figure 1 depicts the steps involved in the process of image classification. Image pre-processing, segmentation, and other methods are typically applied to mammography images.

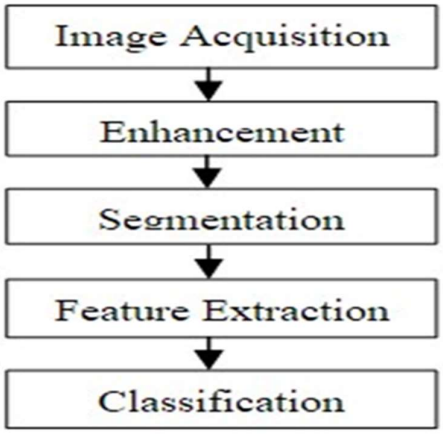


Fig. 1: Image Categorization procedure

This suggested study uses both the KNN and SVM classification techniques to construct a modified version of the SVM classification algorithm that incorporates KNN. In this publication, Section 2 provides a full presentation of the work that is linked. In the third section, a proposed methodology is presented; in the fourth section, experimental results are examined; and in the fifth section, a conclusion is presented.

II Review of Literature

Using neural network training, Marcano-Cedeno et al., 2011 provided a novel approach to enhancing pattern classification. The biological meta plasticity property of neurons and Shannon's information theory inspired the author to suggest a training approach. During the training phase, the Artificial Meta plasticity Multilayer Perceptron (AMMLP) algorithm prioritises updating the weights for the less frequent stimulus over updating the weights for the more frequent stimulus. This method can be used to simulate metaplasticity in a lab. The training efficiency of the AMMLP is optimal, retaining MLP performance. The author used the Wisconsin Breast Cancer Database (WBCD) to put the theory to the test.

Buciu, I., & Gacsadi, A. (2011) propose a solution to the challenge of digital mammography classification. To better distinguish the abnormal regions from what was carefully determined to be the background, the blotches surrounding the tumours were manually eliminated. Gabor wavelets are used to filter the mammography images. Different orientations and frequencies are used to obtain the properties [8]. To achieve the goal of reducing the dimension of filtered and unfiltered high-dimensional data, Principal Component Analysis (PCA) is now being used. Finally, the data was classified using proximal support vector machines. The classification performance of mammogram images was dramatically enhanced by deriving Gabor features as opposed to using the original photographs.

The diagnostic accuracy of micro-calcification has been significantly improved thanks to Wang et al., 2016. The performance of deep learning-based models on huge datasets is analysed and calculated in this paper so that its distinction can be made. A method of semi-automated segmentation is used to characterise each and every one of the micro-calcifications. In order to determine how well micro-calcifications and breast masses can be detected, a distinction classifier model has been developed. Additionally, it can be utilised in either the segregation or integration of breast cancer classification systems. The results were analysed and compared to several benchmark models. When contrasted with the discriminative accuracy produced by their support vector machine model, which was 85.8%, their deep learning model achieved 87.3%.

It has been pointed out that Machine Learning (ML) has developed into an essential component of research in the field of medical image processing (Gardezi SJS, et al., 2019). Over the course of time, an ML approach has progressed from having manually seeded inputs to being able to initialise automatically [9]. The benefits in the field of machine learning have led to a Computer-Aided Diagnosis (CAD) method that is both more intelligent and more independent. The capacity of machine learning algorithms to learn was always expanding and getting better.

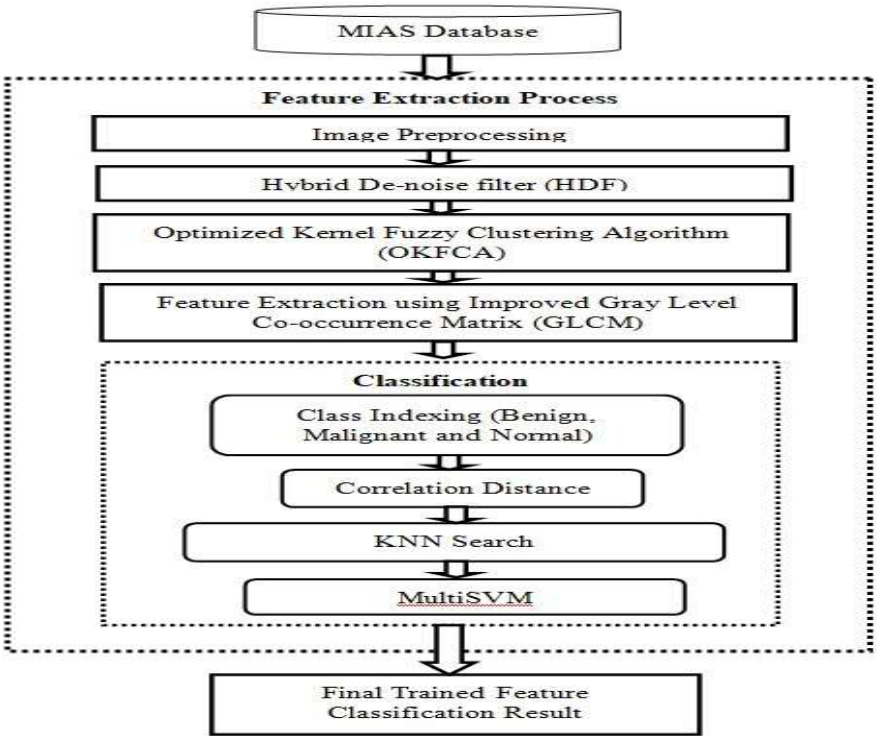
The authors offered a summary of ML and DL approaches for breast cancer that had an exacting purpose.

A HDF algorithm for eliminating noise from mammography images was recently proposed by Dr. D. Devakumari and V. Punithavathi, 2019. This result was accomplished by utilising a combination of the Median Filter and the Applied Median Filter. It was recognised out by utilising Hybrid Denoising Filter (HDF) in order to get rid of the undesirable noises. In this study, a number of different algorithms for de-noising were examined and discussed. The removal of unwanted noise and the length of time required to apply the Hybrid Denoising Filter method were the primary focuses of the experimental findings that were derived from its use of the MATLAB R2013a programme.

An Optimised Kernel Fuzzy Clustering Algorithm was created (OKFCA) and Dr. D. Devakumari and V. Punithavathi, 2020 have recommended using it to locate the cancer sections in mammography pictures. In order to locate the segmented sections contained inside the MIAS database, the OKFCA algorithm has been described. The proposed segmentation algorithm was implemented using pre-processed mammography images, and the proposed OKFCA has been an important method for figuring out whether part of a mammogram contains cancer. The process of data clustering makes it easier to organise data of the same type into a single collection and data of different types into separate groups. The findings of the trials that were performed on the MIAS data provide evidence that the suggested system is effective in terms of accuracy, particularly when contrasted with the performance of the well-known K-Means, OKFCA, and Otsu approaches.

III. METHODOLOGY

The Mammographic Image Analysis Society (MIAS), which is a benchmarked dataset, is used for all of the testing that is performed according to the approach that has been proposed. In this paper, a Hybrid approach that combines a modified SVM with a Multi class KNN classification algorithm is proposed. The KNN and SVM classification models have been expanded in the new approach that has been proposed. The initialization process for this classification begins with the selection of k neighbours, followed by the selection of Training and Test image features. This need for classification is offered for all of the classes that are currently available, and it classifies the breast cancer category using the correlation distance function. Figure 2 depicts the final, fully developed version of the suggested flow diagram.



The proposed flow diagram provides a clear and concise illustration of the process of breast cancer classification. The classification of mammography images begins with class indexing, which is followed by the determination of the correlation distance through the utilisation of KNN and the utilisation of SVM to achieve the classification result. Class indexing is how mammograms are classified.

HDF, OKFCA, and GLCM in Pre-processing Images

The following procedures have already been processed by Dr. D. Devakumari and V. Punithavathi (2019): HDF, image pre-processing, the OKFCA algorithm for segmentation, and the enhanced GLCM feature extraction system are all used in this paper. Lesion detection in mammograms is aided by a process that minimises distracting background noise. Therefore, the Hybrid Denoising Filter (HDF) technique is used to enhance the image quality and refine the precision. Optimised Kernel Fuzzy Clustering Algorithm (OKFCA) is used to segment the cancer sections from the preprocessed image findings displayed in Figs. 3 and 4. This is done after the noise has been removed from the images. The mammograms go through the aforementioned steps before their features are extracted.

Classification Utilising Modified SVM and KNN

A brand new technique known as the Modified SVM and KNN Classification Algorithm is detailed in this article. It is an extension of KNN and SVM that incorporates a Multi class Classification model. These algorithms can be applied to a variety of predictive issues, including classification and regression. This approach to learning is called supervised learning, and it can also be used as a geometric method for classifying data. This classifier gathers all of the existing instances and classifies new instances based on how similar they are to the existing ones (for example, correlation or cityblock distance functions). The KNN approach is utilised during the Feature Extraction process to classify the improved GLCM features that were

extracted and improved. As indicated in Equation 1, the classification was developed utilising correlation or cityblock distance as a measure between the features of the testing data and the reference data. This was done in order to determine which attributes were most similar to one another.

The correlation between vectors A and B is shown as follows in the following representation:

$dist\_Corr$ eqn. (1)

If A and B are two different feature spaces, then the means of A and B are denoted by a and b, respectively. The KNN classification determines estimates of class properties by considering the k training cases that are located closest to each other in the feature space. A dataset for the trained feature extraction of mammography images is provided; this dataset selects the k nearest samples from the classified training data and decides the class by taking into consideration the samples that are most representative of the whole. The user is responsible for determining which option of the constraint k (k N) to employ; this choice is determined by the data of the image. The impact of sound on the classification is mitigated when bigger values are selected for k; however, this results in less clear demarcation between the various classes. Using a variety of various probing approaches, such as cross-validation, it is possible to select a better choice of the value of k. The value of k is chosen to be the one that results in the least amount of classification error.

In machine learning, the SVM (Support Vector Machine) with Multi Class Classification is a strategy that works on the principle of structural risk reduction in order to figure out whether of the three classes are normal, benign, or malignant. The aim of training and testing are served by combining a KNN with an SVM using multi-class data. In the work that is being proposed, the data from the tests are separated into three distinct groups. The data used for the tests were derived from the training data used for the first group. In the second group, the data used for testing were collected separately from the data used for training. The third group's training data were used to generate testing data, and those testing data were obtained both indoors and outside. The purpose of grouping is to assess the level of accuracy achieved by each group. The procedure of categorization is carried out so that mammography pictures can be categorised as either normal or abnormal (benign or malignant).

The SVM that has been upgraded to include KNN has had its functionality expanded to accommodate multiclasses, such as a and b. The value 'a' in the variable indicates a trained feature, while the value 'b' in the variable is a test image feature (Benign, Malignant, or Normal).

In order to acquire a classification process, the proposed method uses as input an OKFCA segmented picture as well as GLCM features. It then proceeds to carry out a number of steps. Figure 3 provides a visual representation of the algorithm that has been suggested.

Algorithm 1: A Hybrid Algorithm with Modified SVM and KNN for Classification

Input: OKFCA Segmented Image, X: training data, Y: class labels of X.

Output: Class *y* (Benign, Malignant and Normal)

Process

Step 1: Choose a value for the parameter *k*.

Step 2: Calculate Modified SVM with KNN

for j = 1 to *m* **do**

Compute distance dist_Corr (A,B) using eqn. (1) **end for**

Step 3: Apply Modified SVM with KNN using eqn. (2)

Step 4: Integrate the classes of these Y samples in one class c.

Step 5: The class of y is c(*Y*) = c

Fig. 3 A Hybrid Algorithm with Modified SVM and KNN for Classification

IV.EXPERIMENTAL RESULTS

The outcome of the experimental has been carried out, and the performance of Modified SVM with KNN Classification has been exhibited. The findings of the experiments were carried out using a computer that had an Intel I5 CPU with a speed of 3.20 GHz 4, 8 gigabytes of random access memory, the Windows 10 operating system, and the MATALB R2013a software. The experimental procedure is carried out using the MIAS picture database. The images are then segmented using OKFCA, and an evaluation of an improved GLCM feature extraction approach is carried out. The MIAS picture database was used in this set of experimental findings, and the Modified SVM with KNN Classification method was used on it. Both the KNN and SVM algorithms that were used were already in existence, and both of these algorithms could be flexibly designed to forecast the exactness or accuracy of the database in order to fulfil the criteria of various tests.

Quantifying the performance of the suggested classification algorithm is often accomplished by aggregating the number of true positives, false positives, and negatives in order to determine the accuracy. They are defined by the following:

$$Modified\ SVM\ with\ KNN\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\ eqn. (3)$$

IV. CONCLUSION

Combining the KNN algorithm with the SVM classification method is what the proposed KNN with SVM Classification algorithm does. This study compares the performance of ten test mammograms with 20 trained mammograms. In addition, the total classification accuracy of 322 mammography pictures taken from the MIAS database is analysed. The results of the experiments show that the classification accuracy for ten test photos using training data is 99.050, which is significantly higher than the existing algorithm's scores of 79.139 and 80.1. The KNN that is being proposed with

SVM shows significantly higher accuracy than KNN for all 322 mammography images in the MIAS database. Due to the intricate nature of its training and classification procedures, the SVM classifier is vulnerable to a systemic statistical issue. This is especially true when there are a big number of features for each training sample and a large number of training samples overall. There are a number of advantages to using hybridised or enhanced models for breast cancer categorization. Using multi-class datasets in training and testing helps achieve this goal, and when combined with conventional algorithmic methods, further boosts accuracy.

REFERENCES

[1] E.C.Fear, P.M.Meaney, and M.A.Stuchly,"Microwaves for breast cancer detection", IEEE potentials, vol.22, pp.1218, February-March 2003.

[2] Homer MJ.Mammographic Interpretation: A practical Approach. McGraw hill, Boston, MA, second edition, 1997.

[3] American college of radiology, Reston VA, Illustrated Breast imaging Reporting and Data system (BI-RADSTM) , third edition, 1998.

[4] P. Shi, S. Ray, Q. Zhu, and M. A Kon. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. BMC Bioinformatics, 12, 2011.

[5] C. Cortes, V. N. Vapnik, "Support vector networks", Machine learning Boston, vol.3, Pg.273-297, September 1995 [6] Girosi f. Jones M. and Poqqio T., "Regularization theory and neural network architectures", Neural computation Cambridge, vol.7, pg.217-269, July 1995.

[7] A. Marcano-Cedeno, J. Quintanilla-Domnguez, and D. Andina. Wbcd breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications, (38), 2011.

[8] Buciu, I. & Gacsadi, A. Directional features for automatic tumor classification of mammogram images. Biomed. Signal Process. Control. 6, 370–378 (2011).

[9] Wang, J. et al. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci.

reports 6, 1–9 (2016).

[10] Gardezi SJS, Elazab A, Lei B, Wang T. Breast cancer detection and diagnosis using mammographic data: Systematic Review. Journal of Medical Internet Research. 2019; 21(7).

[11] Devakumari D and Punithavathi V, "Noise Removal in Breast Cancer using Hybrid De-Noising Filter for Mammogram Images", published by International Conference on Computational Vision and Bio-Inspired Computing (ICCVBIC 2019), January,2020,pp. 109-119.

[12] Punithavathi V and Devakumari D, "A New Proposal for the Segmentation of Breast Lesion in Mammogram Images using Optimized Kernel Fuzzy Clustering Algorithm", accepted by Materials Today: Proceedings - Elsevier, 2020.

[13] Punithavathi V and Devakumari D, "Detection of Breast Lesion Using Improved GLCM Feature Based Extraction

in Mammogram Images", published by SSRN eLibrary, July 30, 2020.