AI Models for Predicting Mammographic Mass Severity

Bukke Devendranaik

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Mammography stands out as the most cost-effective and efficient method for detecting cancer in its preclinical stages, with breast screening programs specifically designed to identify cancer at earlier stages. These screening programs typically yield vast amounts of data, standardized by the Breast Imaging Reporting and Data System (BI-RADS) established by the American College of Radiology. The BI-RADS system provides a standardized vocabulary for radiologists to use when interpreting each finding. The primary objective of this study is to develop AI models that predict mammography outcomes from a reduced set of interpreted mammography findings. However, the low positive predictive value of breast biopsy results stemming from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. In this research paper, data mining classification algorithms, namely Artificial Neural Network (ANN) and Support Vector Machine (SVM), are explored on a mammographic masses dataset. The accuracy of ANN and SVM is reported as 80.3% and 81.9% on test samples, respectively. Our analysis indicates that among these three classification models, SVM predicts the severity of breast cancer with the lowest error rate and highest accuracy.*

I. INTRODUCTION

Breast cancer remains one of the most prevalent diseases among women. In 2016 alone, it is estimated that nearly 246 thousand new cases of invasive breast cancer will be diagnosed, along with 61 thousand cases of non-invasive breast cancer. The journey for any cancer patient, and their caregivers, is undeniably challenging. Early diagnosis of breast cancer is crucial, given its high mortality rate in later stages. Mammography stands as the most reliable method for diagnosing breast cancer in the modern era. The Breast Imaging Reporting and Data System (BI-RADS), established by the American College of Radiology, categorizes mammogram results into four initially, later expanded to six, classifications. Mammography is considered the most cost-effective and efficient technique for identifying risk at a preclinical stage, with breast screening programs specifically aimed at detecting disease in its earlier stages.

The diagnostic evaluation of a patient using the BI-RADS scale may necessitate further biopsy before the physician renders a final diagnosis based on the mammogram. Tumor biopsy outcomes may indicate either malignant or benign tumors. If the tumor is benign, the biopsy could have been avoided, but it is required when the physician is uncertain about a patient's BI-RADS assessment of the mammogram. Nearly 70% of biopsies conducted yield benign results, a significant number of which could have been prevented. Radiologists exhibit considerable variation in interpreting mammography results. In such cases, Fine Needle Aspiration Cytology (FNAC) is employed. However, the average accuracy rate of FNAC identification is only 90%. The goal of BI-RADS identification is to assign a patient to either a benign category, indicating the absence of breast cancer, or a malignant category, suggesting strong evidence of breast cancer. The aim of this study is to enhance the physician's ability to assess the severity of a mammographic mass lesion based on BI-RADS characteristics, thereby reducing unnecessary breast biopsies and considering the patient's age.[1,3,5,7]

II. DATA MINING

Data mining is the process of extracting valuable, previously unknown, and ultimately understandable information from large datasets, using it to make crucial business decisions. The extracted data can be used to form prediction or classification models, or to identify relationships between dataset records. Data mining involves an integration of techniques from various disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, image and signal processing, and spatial or temporal data analysis. By conducting data mining, interesting data patterns, consistencies, or high-level information can be extracted from datasets and viewed or interpreted from various perspectives. The discovered data can be applied to decision-making, process control, information management, and query processing.[2,4]

Data mining is a crucial step in the data discovery process. The main tasks of data mining are generally divided into two categories: Predictive and Descriptive. The goal of predictive tasks is to predict the value of a particular attribute based on the values of other attributes, while for descriptive tasks, the goal is to extract previously unknown and useful information such as patterns, associations, trends, anomalies, and significant structures from large databases. There are several techniques fulfilling

these objectives of data mining. Some of these can be classified into the following categories: clustering, classification, association rule mining, sequential pattern discovery, and analysis.[6]

The development of data mining systems has received much attention in recent years. It plays a vital role in competitive businesses across a wide variety of business environments. It has been widely applied to various tasks such as sales analysis, healthcare, E-commerce, manufacturing, etc. Numerous studies have been conducted on efficient data mining techniques and their relevant applications.

In this research paper, we focused on Classification Rule Mining for data discovery and generated the rules by applying our developed approach to mammographic clinical dataset

III. METHODOLOGY

3.1 Support Vector Machine

The SVM is a new type of machine learning methods based on statistical learning theory. Because of good promotion and a higher accuracy, SVM has become the research focus of the machine learning community. SVMs are set of related supervised learning methods used for classification and regression [9]. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is on the basis of statistical learning theory by Vapnik et al proposed a new learning method, which is built on the basis of a limited number of samples in the information contained in the existing training text to get the best classification results [10].

A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier [9].

3.2 Artificial neural organization (ANN)

An ANN is an information getting ready perspective that is impelled by the way where a characteristic tangible framework in human brain works. ANNs are used comprehensively for the course of action of different issues, including portrayal, vision, talk, plan affirmation, control structures, etc. A colossal number of neurons present in the human frontal cortex structures the vital part of the neural framework perspective and go probably as simple taking care of segments [4]. A fake neuron is a little getting ready unit and plays out a clear computation that is fundamental to the action of a neural framework. The model of a neuron contains the essential parts like wellsprings of data, synaptic burdens, inclination, adding crossing point, and incitation work.

3.2.1 Multilayer Perceptron (MLP)

A MLP is a champion among the most generally perceived Neural Network plan that has been used for various applications. The MLP organize is commonly made out of different centers or dealing with units, and it is figured out into a movement of no less than two layers [6]. The essential layer (or the most diminished layer) is named as an information layer where it gets the external information while the last layer (or the most dumbfounding layer) is a yield layer where the response for the issue is gotten. The disguised layer is the widely appealing layer in the data layer and the yield layer, and may frame with somewhere around one layers. The arrangement of MLP could be communicated as a nonlinear improvement issue. The objective of MLP learning is to find the best loads that limit the differentiation between the information and the yield. The most predominant getting ready estimation used in NN is Back propagation (BP), and it has been used in dealing with various issues in model affirmation and portrayal. This computation depends on a couple of boundaries, for instance, different covered center points at the hid layers learning rate, energy rate, enactment work and the quantity of preparing to happen. Besides, these boundaries could change the exhibition on the gaining from awful to great exactness [2].

3.3 Experimental Results

The analyses have been directed by utilizing Python programming dialect. The Python Scikit-learn is a bundle for information characterization, grouping and representation. We have considered the Mammography mass data from the UCI Machine Learning Repository [8] dataset for experimentation. The Mammography mass data having 961 instances and 6 attributes. In this dataset, 516 instances classified as benign and 445 instances as malignant. There are 162 missing values of different

attributes. The values of ordinal attribute represent categories with some intrinsic ranking while they nominal attribute represent categories with no intrinsic ranking in nominal type.

IV. RESULTS AND DISCUSSION

The whole dataset is divided for training the models and test them by the ratio of 70:30% respectively. The training set is used to estimate each model parameters, while the test set is used to independently assess the individual models.

In this step the mammography dataset has to go through a cleaning process to remove duplicate records and fill missing data. In this data set 162 instances having missing values. The performance of a learning model is dependent on the quality features. Data preparation is an important step when building a model. This phase consists of replace missing data. The proposed stream imputes the missing values then trains and optimizes the two models. So in this step, we replace missing values using Missing imputation strategy as mean was selected. The missing data results are shown in the screen shots of shown in the figure-1 and figure-2.

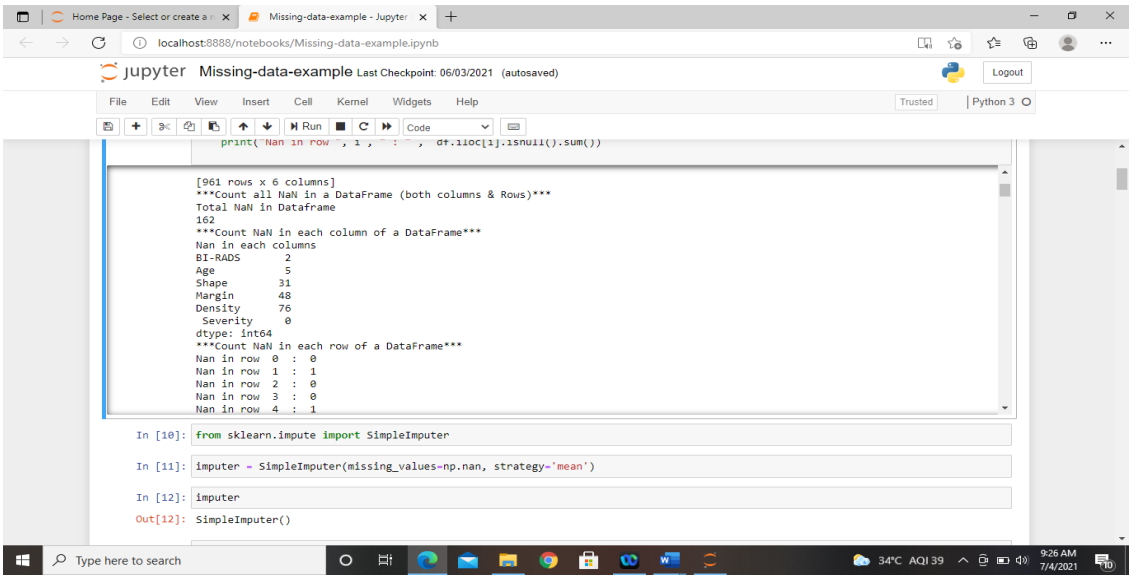

Figure-1: Screen shot of attributes missing records

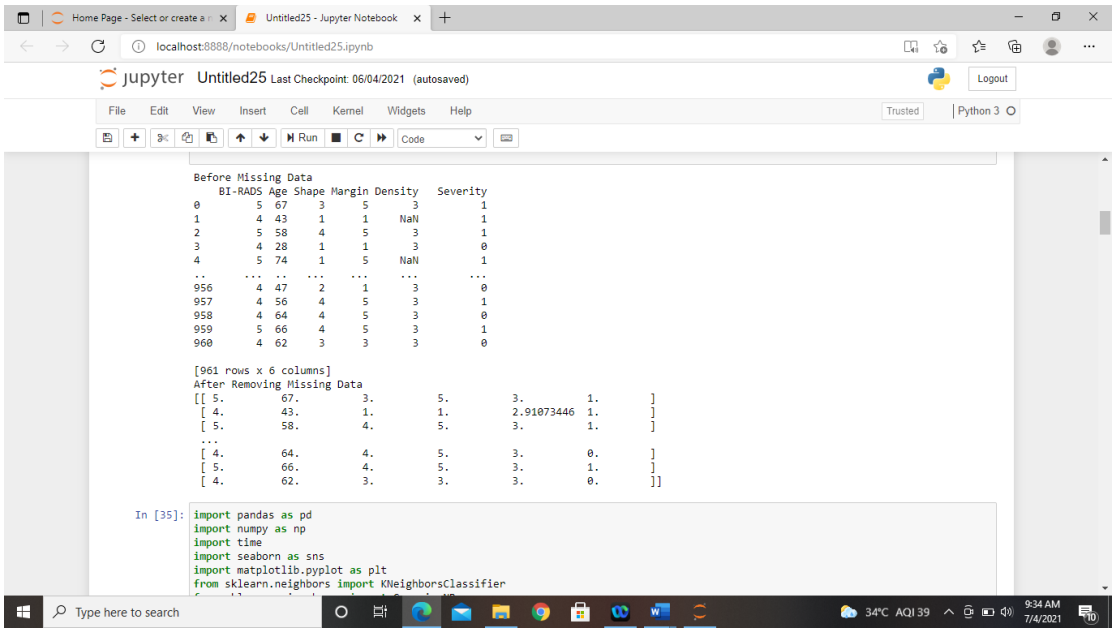

Figure-2: Screen shot of before missing and after filling imputation strategy

In the second stage we implement a SVM and MLP algorithms for prediction of Severity (benign and malignant) of mammographic dataset. The results that we got for MLP and SVM as shown in the figure-3 with their corresponding values.

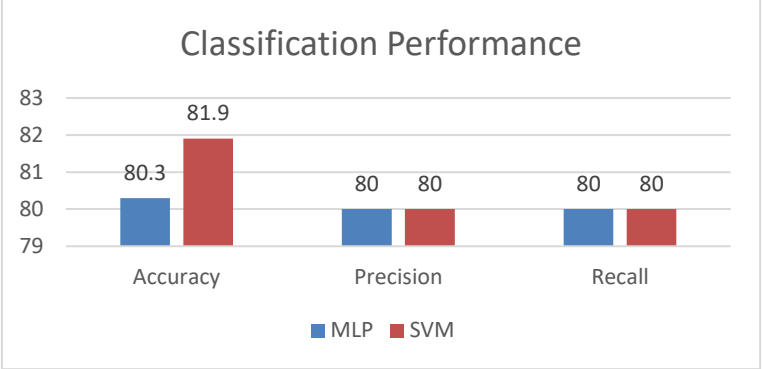

Figure-3: Classification Results

From the figure-3, we observe the performance of MLP accuracy has got 80.3%, whereas the performance of SVM accuracy has achieved 81.9%. However, there is an improvement in the accuracy of SVM over MLP model. The SVM accuracy rate is increased 1.6% over the MLP algorithm. In our experimental result the SVM algorithm shows the highest accuracy compared with MLP. The Experimental screen shots of two models are shown in the figure-4 and figure-5.

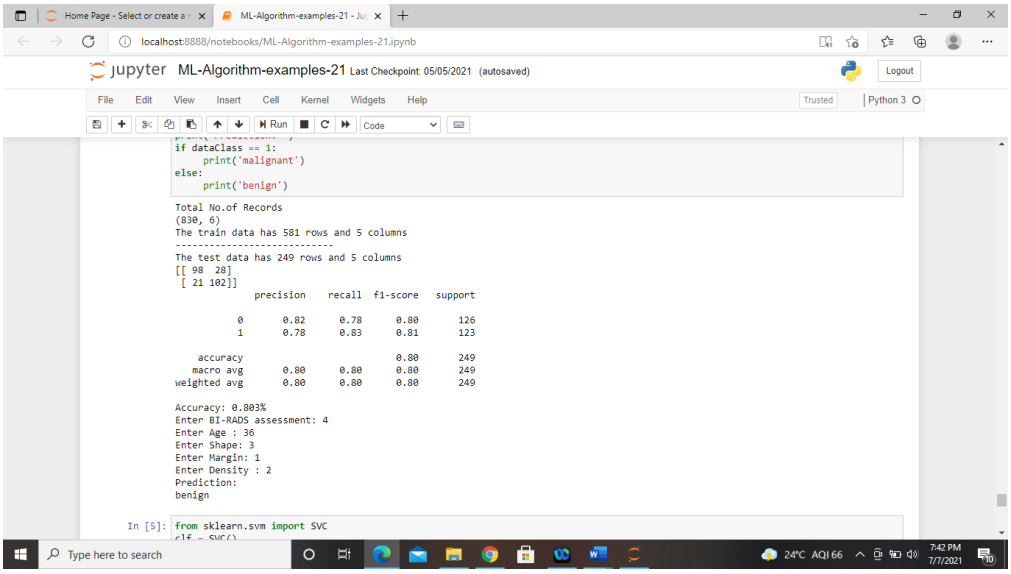

Figure-4: Screen Shot of MLP Algorithm

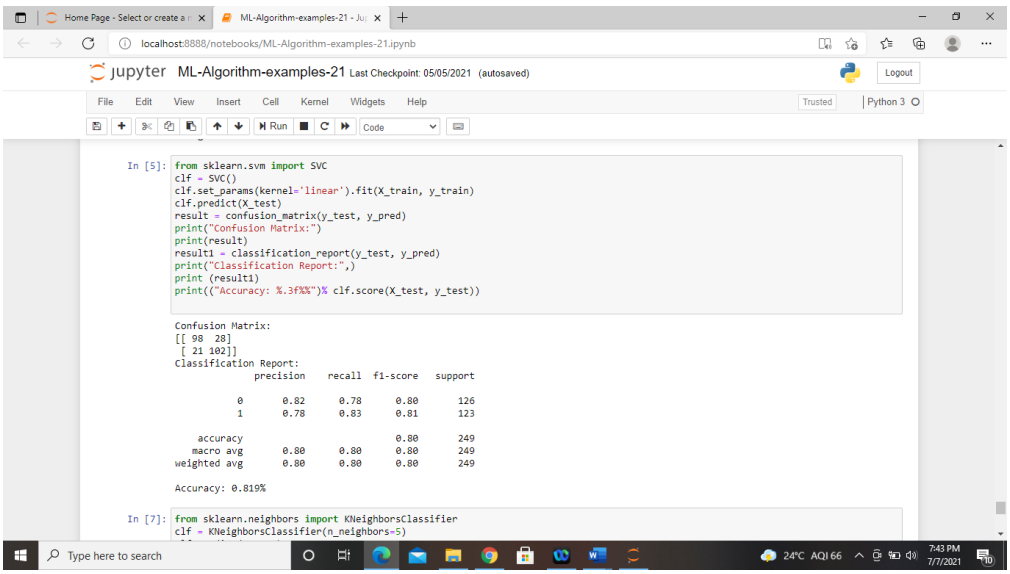

Figure-5: Screen Shot of SVM Algorithm

V. CONCLUSION

In this paper, two different classification models have been analyzed for the prediction of the severity of breast masses. These models are namely artificial neural network and support vector machine. The proposed stream imputes the missing values then trains and optimizes the two models. In this paper mainly focused on to establish an accurate classification model for mammographic mass medical diagnosis. The empirical results reveal that the SVM model does outperform the MLP method in terms of learning accuracy and complexity.

REFERENCES

[1] Elmore, J., M. Wells, M. Carol, H. Lee, D. Howard and A. Feinstein, 1994. Variability in radiologists‟ interpretation of mammograms. N. Engl. J. Med., 331:1493-1499.
[2] H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2006)
[3] http://www.breastcancer.org/symptoms/understand_bc/statistics
[4] J.Han and M.Kamber,”Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.
[5] M. Margaret, Eberl, C.H. Fox, MD, S.B. Edge, C.A. Carter, and M.C. Mahoney, BI-RADS Classification for Management of Abnormal Mammograms, The Journal of the American Board of Family Medicine19, 2006, pp.161-164.
[6] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”, 2nd edition, Addison Wesley, 2005.
[7] Simone A. Ludwig. Prediction of breast cancer biopsy outcomes using a distributed genetic programming approach. In ACM International Health Informatics Symposium, IHI 2010, Arlington, VA, USA, November 11 - 12, 2010, Proceedings, pages 694–699, 2010.
[8] UCI machine learning repository. http://archive.ics.uci.edu/ml/.
[9] Vapnik V.N, “Statistical learning Theory”, John Wiley and Sons, New York, USA, 1998.
[10] Vapnik, V.N. The Natural of Statistical Learning theory. Springer–Verleg,NewYork,USA1995.