IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124113

ENSEMBLE LEARNING FOR BREAST CANCER DIAGNOSIS: RULE CONVERSION AND FEATURE SELECTION APPROACHES WITH MULTIPLE CLASSIFIERS

Mr. Buvaneshraj K[1], Mr. Keerthivasan P[2], Mr. Senthilkumaran S[3], Mr. Veerappan P[4]

Information Technology, Puducherry Technological University, Puducherry, India[1-4]

Abstract: Now-a-days, there are more types of cancers found across all over the world. Somehow, Breast cancer is one of the most common forms of cancer in women worldwide. For a successful treatment of Breast cancer, we need to detect at the earliest stage. There is a challenge in diagnosing breast cancer accurately. Often there is a need for more effective methods to detect and diagnosis the disease. The main objective of this project is to generate rules from the Breast Cancer Wisconsin dataset using a combination of data exploration, feature reduction, and machine learning algorithms. The first step is to explore, pre-process and understand the dataset, followed by rule conversion using random forest algorithm. After the rule conversion, Feature reduction is performed using Extra Tree Classifier, Recursive Feature Elimination (RFE), and correlation between features, and the top 10 features are selected. Then, we used SelectFromModel method for further reduction of reduced features to 5. Again, Rule conversion is performed using random forest algorithm on the selected features. Finally, the generated rules are predicted with the original dataset using several machine learning algorithms such as SVM, MLP, Gradient Booster, Ada Booster, CNN, Extra Tree, and Logistic Regression. By identifying the most important features for predicting breast cancer, we aim to provide clinicians and researchers with valuable insights and tools for more accurate diagnosis and treatment of the disease.

Keywords: Breast cancer, Rule generation, Random Forest, Feature selection, ExtraTreeClassifier, RFE Correlation, SVM, MLP, Gradient boosting, AdaBoost, CNN, Logistic regression

I. INTRODUCTION

Across all over the world, Breast cancer is a leading cause of cancer deaths among women. Detection of breast cancer at their early stage is very crucial for the survival rate of the victims. With the increasing availability of data, machine learning techniques have been used for the prediction of breast cancer.

Rule-based classifiers provide transparency and interpretability, which is essential in the medical domain. In this project, we aim to generate decision rules using machine learning algorithms to predict the diagnosis of breast cancer based on clinical and demographic features.

II. MOTIVATION

The motivation of this project is to develop a machine learning-based approach for breast cancer diagnosis, which can be helpful for doctors and medical professionals in making accurate and timely diagnosis. The main objective of this proposed system is to improve the accuracy and efficiency of the existing diagnostic methods, reduce errors, and provide a reliable and effective tool for early detection and treatment of breast cancer.

III. LITERATURE SURVEY

[1] Conducted a survey on breast cancer classification using association rules and support vector machines (SVMs). Their study focused on reducing the number of features used in breast cancer classification, which can help improve the efficiency and accuracy of classification algorithms. The authors proposed a novel approach that combines association rule mining and SVM classification to select the most informative features for breast cancer diagnosis.

[2] They proposed a new approach that uses machine learning to identify patterns in breast cancer data and extract rules that can be used to make more accurate diagnoses. They used random forest algorithm to classify breast cancer data and identify important features that are predictive of cancer

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023
DOI: 10.17148/IJARCCE.2023.124113

[3] They proposed an approach which combines a neural network with an expert system to identify major risk factors of breast cancer. The authors also incorporate a feature selection algorithm to reduce the dimensionality of the input data and improve the efficiency of the system. The results of their approach outperformed the other methods in terms of accuracy and efficiency, and was able to identify the most significant risk factors associated with breast cancer.

[4] They proposed a novel approach for breast cancer management using decision tree and neural network algorithms. They propose a system that utilizes both decision tree and neural network algorithms for more accurate and efficient breast cancer management. They proposed approach has the potential to improve the accuracy and efficiency of breast cancer management, which can have a significant impact on patient outcomes and healthcare costs

[5] They proposed a novel approach for identifying cancer-related genes using feature selection and association rule mining techniques. Their proposed system first uses a feature selection algorithm to identify a subset of genes that are most relevant to cancer. The authors then use association rule mining to identify frequent patterns of gene expression that are strongly associated with cancer. Then the results of the association rule mining are then used to identify the most significant cancer-related genes.

IV. LMITATIONS IN THE EXISTING SYSTEM

1. In the existing system, Traditional statistical methods are used for breast cancer diagnosis.
2. These methods have limitations such as requiring the data to be normally distributed, which is not always the case in medical datasets.
3. They often assume a linear relationship between features, which may not be true for complex medical datasets like breast cancer.
4. However, traditional machine learning algorithms can be prone to overfitting and are not interpretable.
5. It is difficult to understand how the algorithm arrived at its decision, which can be problematic in a medical context where transparency and interpretability are crucial.
6. Existing studies often focus on only one machine learning algorithm, which may not be the most accurate or effective for a given dataset.

V. PROPOSED SYSTEM

Based on the limitations of the existing system, we propose a new system that addresses these issues and improves the overall performance. The proposed system improves upon the limitations of the existing system by using feature reduction techniques to select the most important features, which in turn improves the accuracy of the generated rules. Additionally, the use of multiple machine learning algorithms for prediction and evaluation provides a more comprehensive understanding of the performance of the system. The proposed system includes the following steps:

- **Data cleaning and pre-processing**: The breast cancer wisconsin dataset is cleaned and pre-processed to remove any missing values, outliers, and other errors that can affect the analysis.

- **Data exploration**: The pre-processed dataset is explored to understand the distribution of the features, their correlations, and other important characteristics.

- **Rule conversion using Random Forest**: The pre-processed dataset is used to generate rules using Random Forest algorithm with 500 trees.

- **Feature reduction**: Several feature reduction methods, including Extra Tree Classifier, Recursive Feature Elimination (RFE), and correlation analysis, are used to select the most important features. The top 10 features are selected, and then reduced to 5 using the SelectFromModel method.

- **Rule conversion for selected features using Random Forest**: The selected features are used to generate rules using the same Random Forest algorithm as step 3.

- **Prediction and evaluation**: The generated rules are used to predict the diagnosis of breast cancer using various machine learning algorithms, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Gradient Boosting, AdaBoost, Extra Tree, Convolutional Neural Network (CNN), and Logistic Regression. The performance of each algorithm is evaluated using metrics such as accuracy, precision, F1-score, and confusion matrix. A comparison chart is generated to compare the performance of each algorithm.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124113

VI. ARCHITECTURE DIAGRAM OF A PROPOSED SYSTEM

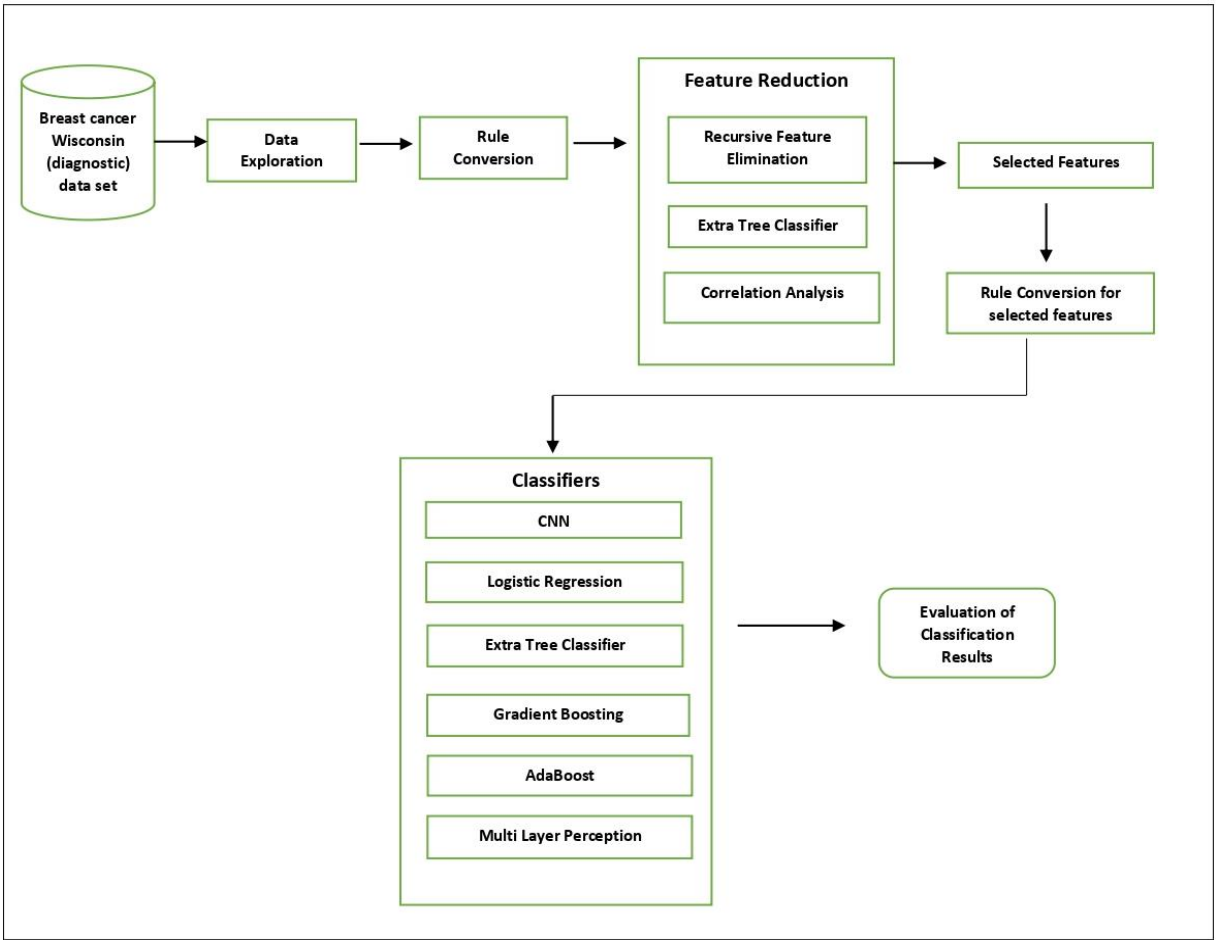

Fig 1. Detailed architecture diagram of proposed system

VII. EXPERIMENTAL RESULTS

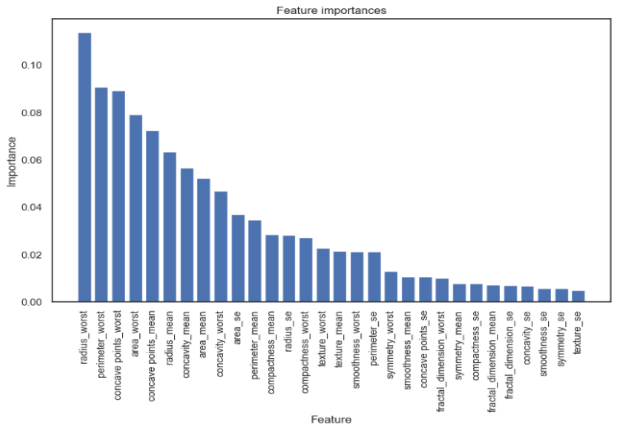

Fig 2. Feature Selection with Extra Tree

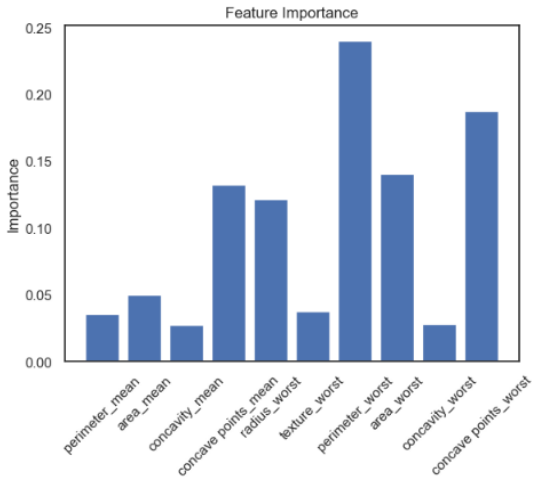

Fig 3. Feature Selection with RFE

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124113

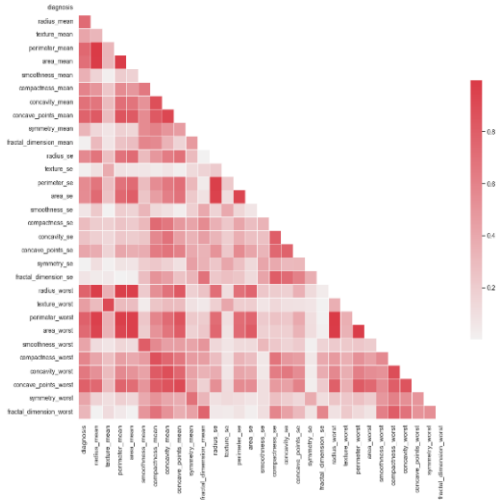

Fig 4. Correlation analysis between features

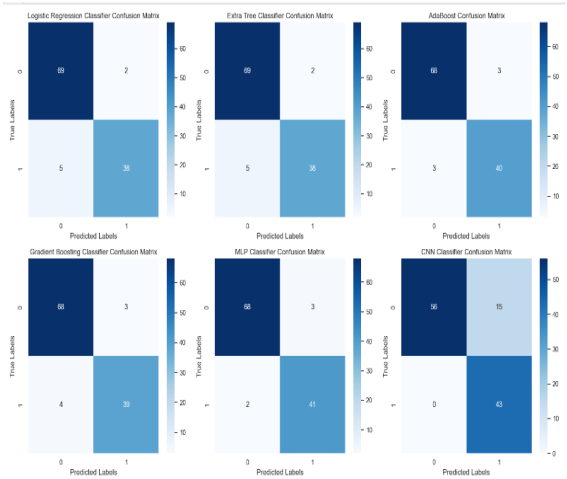

Fig 5. Confusion matrix for classifiers

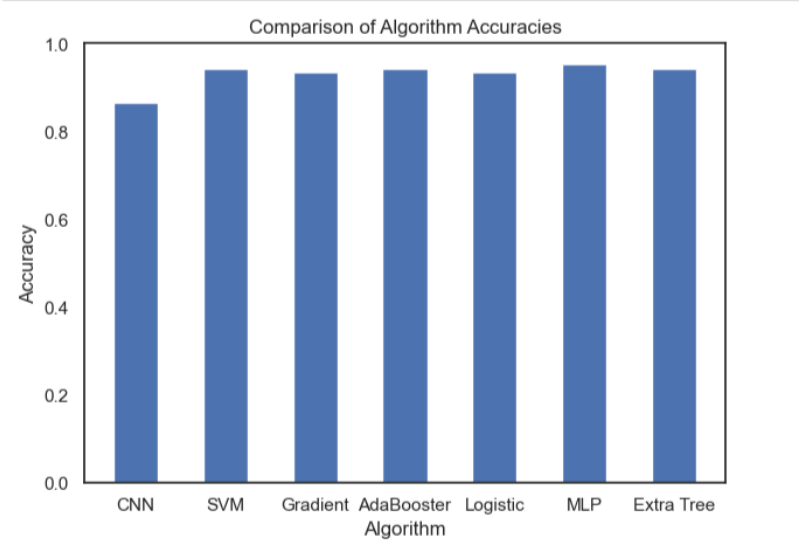

Fig 6. Accuracy Comparison for six classifiers

VIII. CONCLUSION

In conclusion, this project successfully generated rules from the Breast cancer Wisconsin dataset using random forest and feature reduction techniques. The selected features were used to generate new rules, which were then used to predict the diagnosis of breast cancer patients using several machine learning algorithms.

We found that MLP algorithm showing the highest accuracy among the used algorithms. This project demonstrates the potential of using machine learning and rule-based systems for cancer diagnosis and highlights the importance of feature reduction in improving the accuracy and interpretability of the model.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124113

REFERENCES

[1] Ed-daoudy, Abderrahmane, Maalmi, Khalil **"**Breast cancer classification with reduced feature set using association rules and support vector machine" Springer 2020.

[2] Sutong Wang and Yuyan Wang and Dujuan Wang and Yunqiang Yin and Yanzhang Wang and Yaochu Jin **"**An improved random forest-based rule extraction method for breast cancer diagnosis" Elseiver 2020

[3] Das, Akhil Kumar and Biswas, Saroj Kr. and Mandal, Ardhendu and Chakraborty, Manomita "A Neural Expert System to Identify Major Risk Factors of Breast Cancer" IEEE 2020.

[4] Verma, A.K., Chakraborty, M. & Biswas, S.K. "Breast Cancer Management System Using Decision Tree and Neural Network" Springer 2021.

[5] Consolata Gakii, Richard Rimiru "Identification of cancer related genes using feature selection and association rule mining" Elseiver 2021.