Identifying Breast Cancer through the Application of Machine Learning Algorithms: A Comprehensive Exploration of Techniques and Methods for Detection and Diagnosis

G Mallikarjunareddy

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Cancer is a leading cause of mortality globally, with breast cancer posing a significant threat to women's health worldwide. Early detection is key to effective treatment. This study employs machine learning techniques, specifically Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers, to classify breast cancer data. Utilizing SVM-RFE for dimensionality reduction, the study aims to identify the smallest subset of features for improved classification of benign and malignant tumors. By analyzing the Wisconsin Breast Cancer (WBC) dataset, this research seeks to optimize feature selection methods to enhance classification accuracy. Results indicate that MLP classifier achieves higher accuracy rates post feature selection. Comparative analysis of SVM and Artificial Neural Network underscores the efficacy of feature selection techniques in improving classification performance.*

I. INTRODUCTION

The advancement of digital diagnostics has been driven by the imperative to support medical professionals in decision-making. Initially applied in healthcare for tasks like electrocardiograms, their usage has expanded to encompass ultrasounds and other traditional diagnostic methods. Traditionally, disease detection and monitoring heavily relied on healthcare professionals. However, the growing number of patients requiring continuous assessment has propelled technical advancements in automated systems. Converting qualitative data into quantitative measures is pivotal in addressing diagnostic challenges. Cancer, a broad term encompassing a multitude of diseases affecting various body parts, is characterized by the rapid proliferation of abnormal cells that surpass their normal boundaries, infiltrating adjacent tissues and potentially spreading to other organs—a process known as metastasis. Metastases are the primary cause of cancer-related deaths worldwide. Breast cancer, specifically, ranks as the second leading cause of cancer mortality among women aged 40 to 55, with an estimated 1.2 million new cases diagnosed annually according to projections by the World Health Organization.

Cancer is characterized by uncontrolled cell growth within the body, often named after the affected body structure. Breast cancer, notorious for its high mortality rate in women, manifests as rapidly dividing cells forming masses within the breast—referred to as tumors. Tumors are categorized as benign or malignant, with malignant tumors invading healthy tissues and potentially spreading to other organs, causing further damage. Breast cancer specifically denotes a malignant tumor within the breast.

Consequently, numerous studies have focused on early cancer detection, given its detrimental impact on human health. This study aims to diagnose cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset."[1][2][8][9].

II. FEATURE SELECTION

Highlight determination issue is maybe the primary issues in data portrayal. The inspiration driving component choice will be decision of the most un-number of highlights to grow exactness and lessening the cost of data gathering [4]. Of late, in view of appearance of high-dimensional datasets with low number of tests, plan models have encountered over-fitting issue. Along these lines, the necessity for include choice systems that are used to dispense with the developments and unimportant highlights is felt [5][7].

For precise estimate incorporate assurance is huge. Data mining estimations used part decision techniques for picking the best highlights from the dataset. These features or characteristics should be stacked directly into the memory for preprocessing. Highlight assurance is a collaboration where simply the subset of the reasonable highlights is picked [14]. This procedure recognizes two or three most critical attributes and help to expect the outcome. It's anything but a kind of dimensionality decline used for preprocessing. The qualification between incorporate decision and dimensionality decline is the main method (Feature decision) will decrease the characteristics without making change in the instructive file [3][4][5]. Since incorporate decision strategy oversees less limit it will decrease the multifaceted nature. There are various procedures for incorporate decision estimations applied in portrayal. They are I) Filter method ii) Wrapper Technique and iii) Embedded methodology [7][10]. The channel strategies are used to pick the features subject to the scores in various verifiable connections. Covering strategy uses an energetic approach in incorporate decision. It surveys all possible mix and conveys the outcome for Machine learning. The introduced strategy merges the advantage of two models.

2.1 Support Vector Machine-Recursive Feature Elimination (SVM-RFE)

The especially considered SVM-RFE estimation [7] is a covering feature decision strategy which makes the situating of features using backward component end. It was at first proposed to perform quality assurance for sickness request [4]. Its essential idea is to take out dull characteristics and yields better and more modest quality subsets. The features are cleared out as demonstrated by a premise related to their assistance to the detachment work, and the SVM [15] is re-arranged at every movement. SVM-RFE is a weight-based technique; at every movement, the coefficients of the weight vector of a direct SVM are used as the component situating model [4].

The SVM-RFE computation [6] can be broken into four phases:

1. Train a SVM on the arrangement set;
2. Solicitation features using the heaps of the resulting classifier;
3. Discard features with the tiniest weight;
4. Repeat the connection with the arrangement set restricted to the extra features.

III. METHODOLOGY

This section gives the concise thought of chosen administered models of Support Vector Machine and Multilayer Perceptron.

3.1 Support Vector Machine (SVM)

SVM was presented by Vapnik and it's anything but a strategy dependent on the factual learning hypothesis and has been applied for tackling order and relapse issues [12]. The target of the SVM is to isolate two classes by deciding the direct classifier that amplifies the edge and it is alluded to as the ideal isolating hyperplane [13]. SVM has been utilized in different order issue and generally current interest in bosom malignancy discovery due its heartiness. The regularization boundary and portion work are the two significant segments that need to been resolved prior to directing preparing. A portion of the critical explores utilized utilizing the SVM for bosom malignant growth location used heuristics SVM approaches, for example, the smooth SVM, the direct SVM and general non straight SVM [12]. The objective of SVM is to decide a reasonable hyperplane with most extreme edge which can be processed as an advancement issue [10].

3.2 Multilayer Perceptron (MLP)

A MLP is a hero among the most by and large saw Neural Network plan that has been utilized for different applications. The MLP put together is generally made out of various focuses or managing units, and it is sorted out into a development of no under two layers [7]. The fundamental layer (or the most reduced layer) is named as a data layer where it gets the outer data while the last layer (or the most stunning layer) is a yield layer where the reaction for the issue is gotten. The hidden layer is the broadly engaging layer in the information layer and the yield layer, and may outline with somewhere near one layers. The course of action of MLP could be imparted as a nonlinear improvement issue. The goal of MLP learning is to track down the best loads that limit the separation between the data and the yield. The most prevalent preparing assessment utilized in NN is Back spread (BP), and it has been utilized in managing different issues in model attestation and depiction. This calculation relies two or three limits, for example, unique covered focus focuses at the concealed layers learning rate, energy rate, order work and the amount of getting ready to occur. Additionally, these limits could change the show on the acquiring from dreadful to incredible precision.

IV. EXPERIMENTAL RESULTS

The investigations have been coordinated by using Python programming tongue. The Python Scikit-learn is a pack for data portrayal, gathering and portrayal. The Breast Cancer Wisconsin (Diagnostic) dataset utilized in this examination was taken from the Irvine Machine Learning Repository of the University of California (UCI) [11]. The Breast Cancer Wisconsin (Diagnostic) informational collection has 569 lines and 32 sections. This information contains two class marks i.e., The Benign class has 357 occurrences and harmful class cases has 212. To approve the forecast aftereffects of the examination of the two order (SVM and MLP with SVM-RFE) strategies and the 10-overlap hybrid approval is utilized. The k-crease hybrid approval is generally used to diminish the blunder came about because of arbitrary examining in the correlation of the exactnesses of various forecast models. We utilize 70% of records as the preparation information and the other 30% as the testing information. Order exactness (%) rates got without Feature Selection and with Feature Selection for two unique systems (MLP and SVM) have been displayed in the Figure-1 and same displayed in the table-1.

Table-1
Performance of the two models

Algorithm	Accuracy	Precision	Recall
MLP	93.56	94	94
MLP with selected features	96.49	96	96
SVM	91.81	92	92
SVM with selected features	94.15	94	94

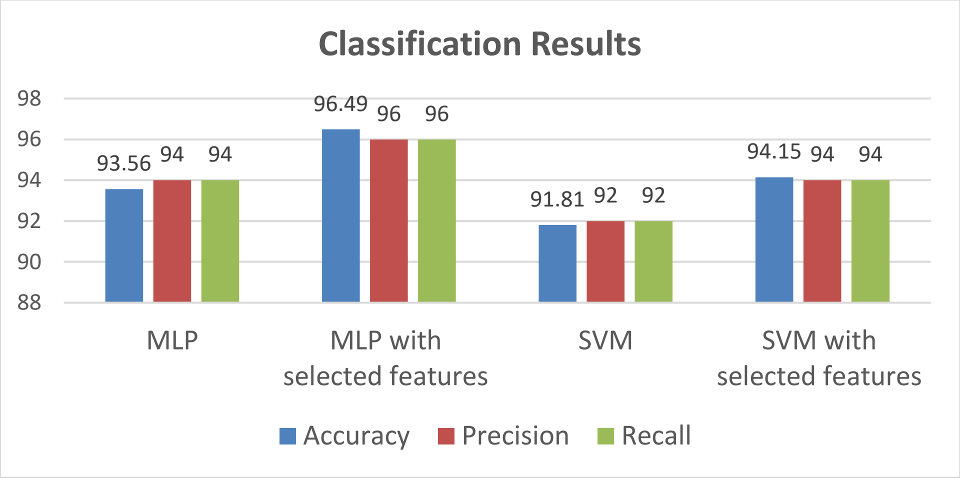
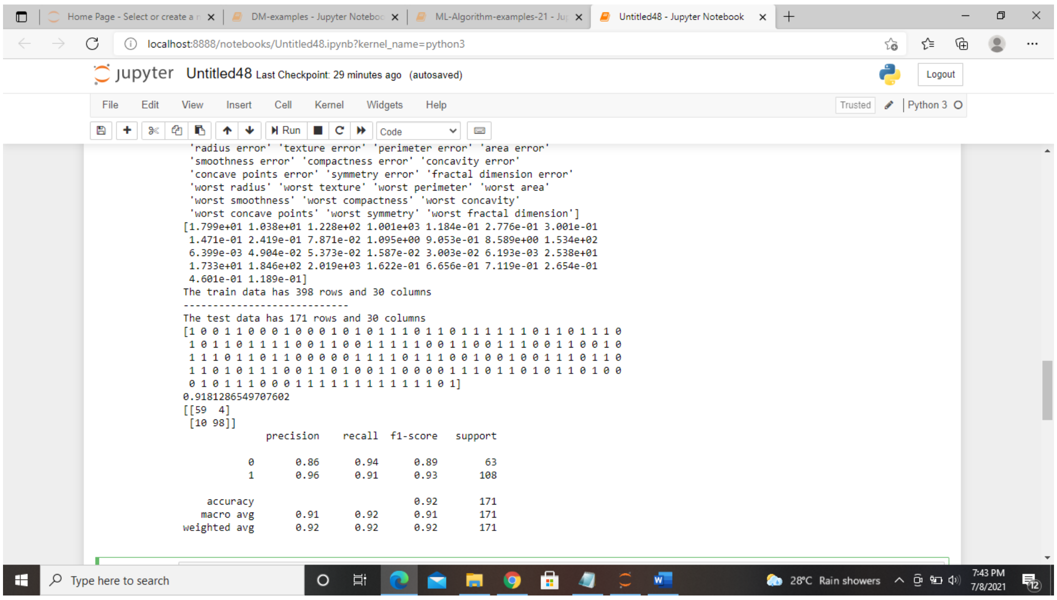


Figure-1: Performance of the two models

We see in the figure-1, the presentation of the two order calculations with SVM-RFE based component determination and without highlight choice on the dataset. The accuracy of MLP calculation has accomplished 93.56% while MLP with SVM-RFE 96.49%. The accuracy of SVM calculation without SVM-RFE has 91.81%, while utilizing SVM with SVM-RFE has 94.15%.

The detailed experimental result screen shots are shown from the figure-2 to figure-5.

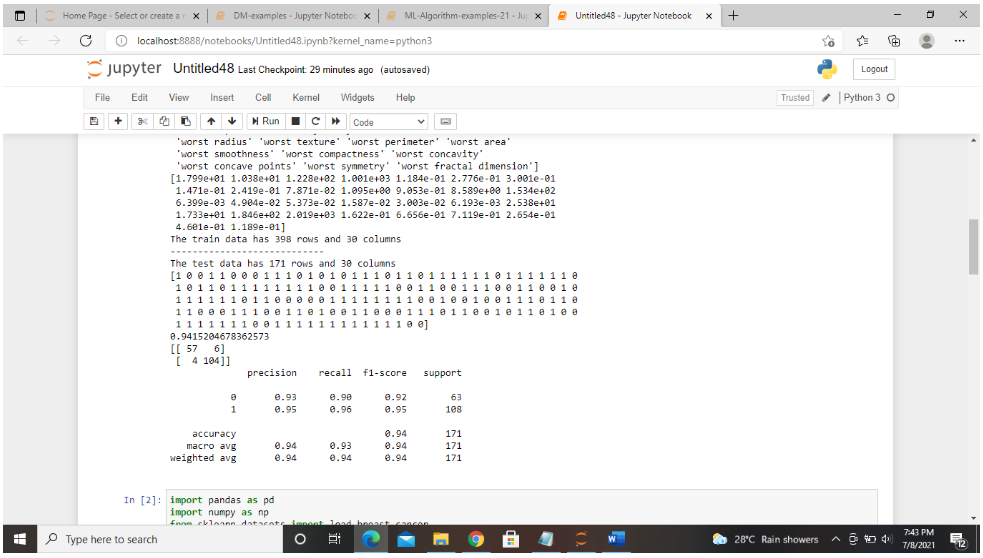

Home Page - Select or create a n
DM-examples - Jupyter Noteboo
ML-Algorithm-examples-21 - Ju
Untitled48 - Jupyter Notebook
localhost:8888/notebooks/Untitled48.ipynb?kernel_name=python3
jupyter
Untitled48 Last Checkpoint: 29 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Type here to search
28°C Rain showers
7:43 PM
7/8/2021

```
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
[1.799e+01 1.038e+01 1.228e+02 1.001e+03 1.184e-01 2.776e-01 3.001e-01
 1.471e-01 2.419e-01 7.871e-02 1.095e+00 9.053e-01 8.589e+00 1.534e+02
 6.399e-03 4.904e-02 5.373e-02 1.587e-02 3.003e-02 6.193e-03 2.538e+01
 1.733e+01 1.846e+02 2.019e+03 1.622e-01 6.656e-01 7.119e-01 2.654e-01
 4.601e-01 1.189e-01]
The train data has 398 rows and 30 columns
----------------------------
The test data has 171 rows and 30 columns
[1 0 0 1 1 0 0 0 1 0 0 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 0 1 1 1 0
 1 0 1 1 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0
 1 1 1 0 1 1 0 1 1 0 0 0 0 0 1 1 1 1 0 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 0 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 0 0
 0 1 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1]
0.9181286549707602
[[59  4]
 [10 98]]
              precision    recall  f1-score   support

           0       0.86      0.94      0.89        63
           1       0.96      0.91      0.93       108

    accuracy                           0.92       171
   macro avg       0.91      0.92      0.91       171
weighted avg       0.92      0.92      0.92       171
```

Figure-2: Results of SVM



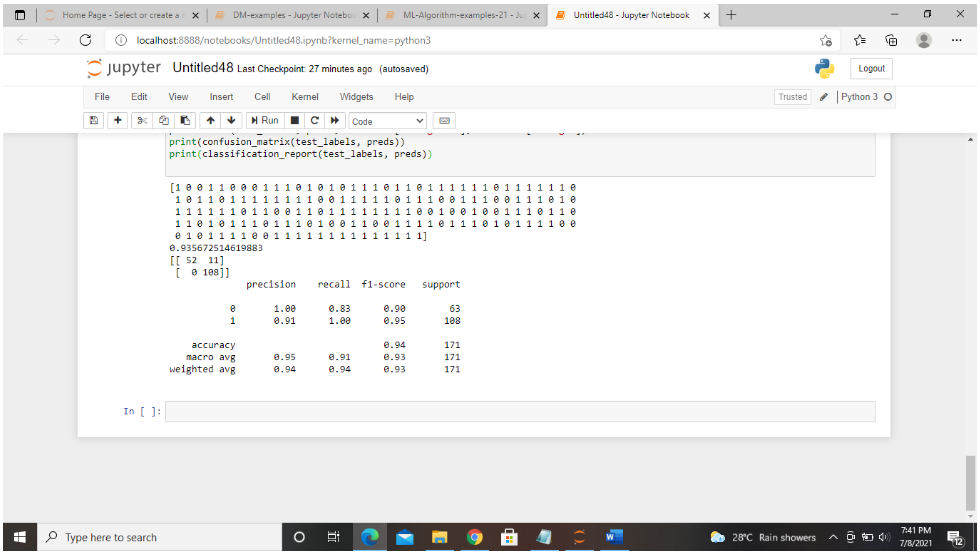
```
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
[1.799e+01 1.038e+01 1.228e+02 1.001e+03 1.184e-01 2.776e-01 3.001e-01
 1.471e-01 2.419e-01 7.871e-02 1.095e+00 9.053e-01 8.589e+00 1.534e+02
 6.399e-03 4.904e-02 5.373e-02 1.587e-02 3.003e-02 6.193e-03 2.538e+01
 1.733e+01 1.846e+02 2.019e+03 1.622e-01 6.656e-01 7.119e-01 2.654e-01
 4.601e-01 1.189e-01]
The train data has 398 rows and 30 columns
----------------------------
The test data has 171 rows and 30 columns
[1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0
 1 1 1 1 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 0 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 1 0 1 1 0 0 1 0 1 1 0 1 0 0
 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0]
0.9415204678362573
[[ 57   6]
 [  4 104]]
              precision    recall  f1-score   support

           0       0.93      0.90      0.92        63
           1       0.95      0.96      0.95       108

    accuracy                           0.94       171
   macro avg       0.94      0.93      0.94       171
weighted avg       0.94      0.94      0.94       171
```

```
In [2]: import pandas as pd
        import numpy as np
        from sklearn.datasets import load_breast_cancer
```

Figure-3: Results of SVM with selected features



```
print(confusion_matrix(test_labels, preds))
print(classification_report(test_labels, preds))
```

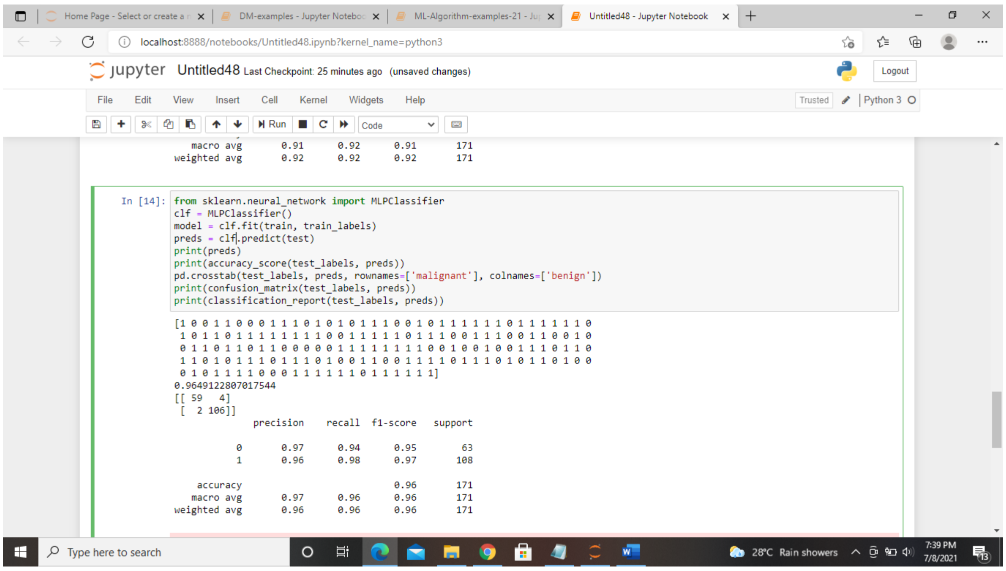
```
[1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 1 0
 1 1 1 1 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 0
 0 1 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
0.935672514619883
[[ 52  11]
 [  0 108]]
              precision    recall  f1-score   support

           0       1.00      0.83      0.90        63
           1       0.91      1.00      0.95       108

    accuracy                           0.94       171
   macro avg       0.95      0.91      0.93       171
weighted avg       0.94      0.94      0.93       171
```

```
In [ ]:
```

Figure-4: Results of MLP



```
   macro avg       0.91      0.92      0.91       171
weighted avg       0.92      0.92      0.92       171
```

```
In [14]: from sklearn.neural_network import MLPClassifier
         clf = MLPClassifier()
         model = clf.fit(train, train_labels)
         preds = clf.predict(test)
         print(preds)
         print(accuracy_score(test_labels, preds))
         pd.crosstab(test_labels, preds, rownames=['malignant'], colnames=['benign'])
         print(confusion_matrix(test_labels, preds))
         print(classification_report(test_labels, preds))
```

```
[1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0
 0 1 1 0 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 0
 0 1 0 1 1 1 1 0 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1]
0.9649122807017544
[[ 59   4]
 [  2 106]]
              precision    recall  f1-score   support

           0       0.97      0.94      0.95        63
           1       0.96      0.98      0.97       108

    accuracy                           0.96       171
   macro avg       0.97      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```

Figure-5: Results of MLP with selected features

So, in these two datasets, SVM and MLP calculations with SVM-RFE highlight determination has hot most noteworthy correct nesses when contrasted with just SVM and MLP order.

V. CONCLUSION

The precise early detection of breast cancer cells can be facilitated through the utilization of AI techniques, potentially reducing healthcare costs and expediting treatment initiation for patients. This paper explores the application of Multilayer Perceptron (MLP) and Support Vector Machine (SVM) algorithms for diagnostic and predictive assessment of breast cancer. Consequently, feature selection methods have become imperative in several studies. In this research, a comparative analysis was conducted based on SVM-RFE-based feature selection algorithms to predict the risks associated with Breast Cancer Wisconsin (Diagnostic) infection. We proposed an SVM-RFE based feature selection method for classification tasks, aiming to integrate SVM-RFE computation with MLP and SVM algorithms to enhance classifier accuracy. Our experimental results indicate that reusing features eliminated during the SVM-RFE process can improve the performance of the MLP classifier. MLP is found to outperform SVM by providing higher prediction accuracy."

REFERENCES

[1] Akay, M., “Support vector machines combined with feature selection for breast cancer diagnosis”, Expert systems with applications, Vol.36, 2009, pp.3240-3247

[2] C. Fitzmaurice, C. Allen, and R. Barber, "A systematic analysis for the Global Burden of Disease Study," JAMA Oncol, vol. 3, pp. 524-548, 2017.

[3] G. Ravi Kumar, K. Nagamani and G. Anjan Babu, “A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction”, Lecture Notes on Data Engineering and Communications Technologies, ISBN 978-981-15-0977-3, Volume 37, PP:173-180, Springer Nature Singapore Pte Ltd. 2020

[4] Guyon, Weston, Barnhill, and Vapnik, “Gene selection for cancer classification using support vector machines,” MACHLEARN: Machine Learning, vol. 46, (2002).

[5] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering”, IEEE Trans. Knowl. Data Eng, vol. 17, no. 4, (2005), pp. 491–502

[6] H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2000)

[7] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, (2003) March, pp. 1157–1182

[8] Mu, T., and Nandi, A., “Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier”, Journal of the Franklin Institute, Vol. 344, 2007, pp.285-311.

[9] O. WH. (2018, 10.01.2018). Cancer. Available: http://www.who.int/en/news-room/fact-sheets/detail/cancer

[10] 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.

[11] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/.

[12] V. N. Vapnik, “The nature of statistical learning theory”, New York, NY, USA: Springer-Verlag New York, Inc., (1995).

[13] Vapnik V.N, “Statistical learning Theory”, John Wiley and Sons, New York, USA, 1998.

[14] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," Journal of Biomedical Informatics, vol. 43, pp. 15-23, 2010.