Semi-supervised Learning - An Alternative to Traditional Breast Cancer Prediction

Suthari Manikanta
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amaravati, India
av.en.u4aie22060@av.students.amrita.edu

Shaik Mohammed Rasheed
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amaravati, India
av.en.u4aie22053@av.students.amrita.edu

Jonnala Kowshik Reddy
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amaravati, India
av.en.u4aie22014@av.students.amrita.edu

Seemakurthi Naga Surya Bhavani Sankar
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amaravati, India
av.en.u4aie22058@av.students.amrita.edu

Avula Venkata Sai Harshendra
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amaravati, India
av.en.u4aie22059@av.students.amrita.edu

Dr.M.Srinivas
Dept. of Computer Science,
Amrita Vishwa Vidyapeetham,
Amaravati, India
m_srinivas@av.amrita.edu

***Abstract*—Breast cancer continues to create diagnostic challenges due to its overlapping and diverse characteristics, complicating the diagnosis. This study evaluates various cases for each stage of preprocessing and optimization techniques in Supervised Learning (SL) and Semi-Supervised Learning (SSL) to attain optimal predictive performance. Nine machine learning classifiers were utilized for both SL and SSL Models for training and testing on the Wisconsin Diagnostic Cancer Dataset using the following algorithms: 1) Logistic Regression (LR), 2) Gaussian Naive Bayes (GNB), 3) Linear-SVM, 4) RBF-SVM, 5) DT, 6) RF, 7) XGBoost, 8) Gradient Boosting (GB), and 9) K-Nearest Neighbors (KNN).This process incorporated feature extraction techniques, including Linear Discriminant Analysis (LDA), Feature Agglomeration, and feature selection through Regularization, employing both L2 and Bayesian methods. For hyperparameter optimization, a comprehensive approach Stratified K-fold and nested cross-validation was adopted and the Model has performed well by acheiving an accuracy of 99% compare to SL, which has showcased that SSL can be an alternative approach for sl when the labeled data is expensive or high computational resources are required.**

***Index Terms*—Breast Cancer Diagnosis, Semi-Supervised Learning, Supervised Learning, Feature Extraction, Hyperparameter Optimization**

I. Introduction

Breast Cancer is among the highest often diagnosed forms of cancers and stands as fifth leading cause of morality due to cancer. Stats provided by GLOBOCAN 2020 states that, around 2.3M new cases of Breast Cancer are currently being diagnosed across the globe. Aside from being the most prevalent form of cancer, breast cancer stands as leading cause of death from cancer among women all over the globe. According to forecasts, the death rate from breast cancer might rise to 61.7% by the year 2040 in the Southeast Asia area.In addition to the fact that breast cancer was responsible for 25% of all instances of cancer that occurred in females.

In India, breast cancer accounts for 28.2% of all cancers diagnosed in women, and it is estimated that there will be 216,108 cases reported in the year 2022 [1] . The most prevalent characteristic is the presence of thick breast tissue, which is a primary factor contributing to the development of breast cancer. According to the National Centre for Disease Informatics and Research, the most concerning aspect is that only 15% of breast cancers in India are detected at the first stage, whereas in developed countries, this percentage is as high as 50%. Although approximately 48% of women are diagnosed before the age of 50, this disparity in early detection has a significant impact on treatment outcomes and survival rates. This indicates that the majority of people who are affected in India are affected at a young age between 1990 and 2016.The age-based breast cancer incidence rate in females rose by 39.1% , a trend that has been observed on a consistent basis across all Indian states over the course of the past 26 years [2].

Breast cancer is associated with many risk factors the primary ones being obesity, insufficient physical activity, alcohol use, and advanced age. Insufficient implementation of therapeutic and preventive measures may lead to the proliferation of tumors throughout the body, therefore creating a very perilous scenario.MRI and microscopic analysis of tumor activity are often used techniques to describe the specific characteristics of cancer, irrespective of its benign or malignant nature [3]

A malignant tumor is a potentially fatal kind of cancer that impacts nearby tissues and can lead to metastatic in nature. On the other hand, a benign tumor is better managed with well-defined boundaries and is less likely to provoke life-threatening complications. Accordingly, in India, the doctor-to-patient ratio is well below the recommended of the World Health Organisation (WHO). Therefore, it is more advantageous to enhance the performance of breast cancer diagnostics by using contemporary methodologies that use machine-learning algorithms. In order to facilitate early identification and risk assessment, computational methods use past clinical records along with corresponding patient data to construct

prediction models .

This study presents a comprehensive examination of SL and SSL algorithms for prediction of breast cancer. The analysis specifically concentrates on several preprocessing techniques that have the potential to decrease data availability, reduce costs related to data annotation, simplify the prediction process, and provide reliable and precise predictions for complex datasets. We assessed the effectiveness of the suggested technique by examining 9 Machine learning models, including LR, DT, RF, voting classifiers, and XGBoost, to determine their performance for the SL and SSL methods. We provide a comprehensive analysis of both SL and SSL methods comparing assessment criteria to assess the effectiveness of both algorithms for improved breast cancer prediction.

II. Literature Review

P. Singh investigated SSL methods to enhance diagnostic precision. The Farthest-First Clustering approach generates new cluster centroids based on their greatest distance from current centroids to enhance variety in data representation. This method was used with ML techniques, including RF and DT, which attained accuracies of 98% and 95.2%, respectively [4]

Shanbehzadeh further used the correlation-based feature selection (CFS) technique, which identifies features according to their reliance on the target variable and their independence from other characteristics. The models were then used with confidence-weighted voting (CWV), an ensemble learning method that integrates confidence ratings into the decision-making process. The amalgamation of these techniques led to Random Forest with Wrapper-J48 with an accuracy of 78.8689% after feature selection (FS) [5] .

K. Rustagi developed a method that combines Salp Swarm Optimization with gray Wolf Optimization, inspired by natural behaviors—salps create chain-like formations, whilst gray wolves display a sophisticated social hierarchy in their hunting tactics. This foundation allowed the use of SVM and KNN algorithms with Confidence Voting, with a remarkable accuracy of 99.42% [6].

S. Prustya devised an approach that improves machine learning decision-making by integrating fuzzy logic with an Interval type-2 membership function, in conjunction with feature extraction methods known as MLF-DR. This method, specifically using the LR-PCA methodology, achieved an acc of 99.1%, with 97.7% sensitivity, 100% specificity, and an F1-score of 98.8% [7]

A. Batool introduced an innovative ensemble architecture that integrates a voting classifier with four unique ML models: Light-GB Machine, LDA, ExtraTrees and Ridhe Classifiers. This framework exhibited a testing accuracy of 97.6%, precision of 96.46%, F1 score of 98.1% and an recall of 100%, thereby mitigating class imbalance and improving classification performance [1] .

K. Parashar suggested a methodology using Support Vector Machines (SVM) and the Rainfall Optimization Algorithm (RFOA), a meta-heuristic inspired by the descent of raindrops, with the objective of enhancing the accuracy of SVM binary classification by parameter optimization [8]

I. Singh devised a model that amalgamates the Artificial Bee Colony (ABC) and Black Hole (BH) algorithms with genetic operators, termed GBHABC, using Support Vector Machine (SVM) to get an accuracy rate of 99.42% on UCI dataset [4].

V. Nemade used many ML approaches for breast cancer prediction, including SVM, KNN, XGBoost, NB, DT, LR, RF, and AdaBoost on a breast cancer dataset. The decision tree and XG Boost classifier exhibited the best accuracy of 97%, with XG Boost also attaining the highest AUC value of 0.999 [9].

III. Methodology

A. Dataset Description

We have taken two datasets one is Wisconsin(diagnostic) (WDBC) and other is Wisconsin (original) (WBCD) this is a publicly available open source datasets, which were obtained from University of California, Irvine ML Repo. It was contributed by Dr. William H. Wolberg, who gathered the data from the University of Wisconsin-Madison's affiliated hospitals between 1989 and 1991.

The experiments were first conducted on Diagnostic wisconsin breast cancer dataset, where the dataset is consisting 569 instances with 30 attributes which were extracted by computing digitial images of breast mass where features includes radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension represented with three categories: 1) average 2) std 3)loss. The data distribution of 357 Benign and 212 Malignant records out of 569. For validation we have utilized Wisconsin Original dataset which cosisits of 699 instances with 9 features. which were consisting 241 Malignant and 458 Benign records. Features consisted like size,smoothness and concavity of the cell nuclei. This repos are widely used in Machine-Learning education & research, particularly to demonstrate classification algorithms and techniques in medical diagnosis. [10] [10]

B. Data Preprocessing

Data preprocessing is crucial step for improving the models performance. The missing values present in the dataset were replaced with median values.Presence of outliers can defect the results and can mislead to the conclusions(misclassifications) therfore they were detected and removed using the Z score technique.Scaling is crucial step to ensure that all features contribute equally in correct predictions, so if features were not properly scaled, those with larger ranges will disproportionately affect distance calculations, resulting in biased outcomes thus scaling process is an important step in preprocessing stage and it helps in aligning the data within a range such that model's performance can be improved in predicting the presence or absence of cancer.

C. Supervised Learning

Supervised learning, a traditional method in machine learning, commences with the selection and analysis of a dataset centered on the target variable. A regression difficulty arises

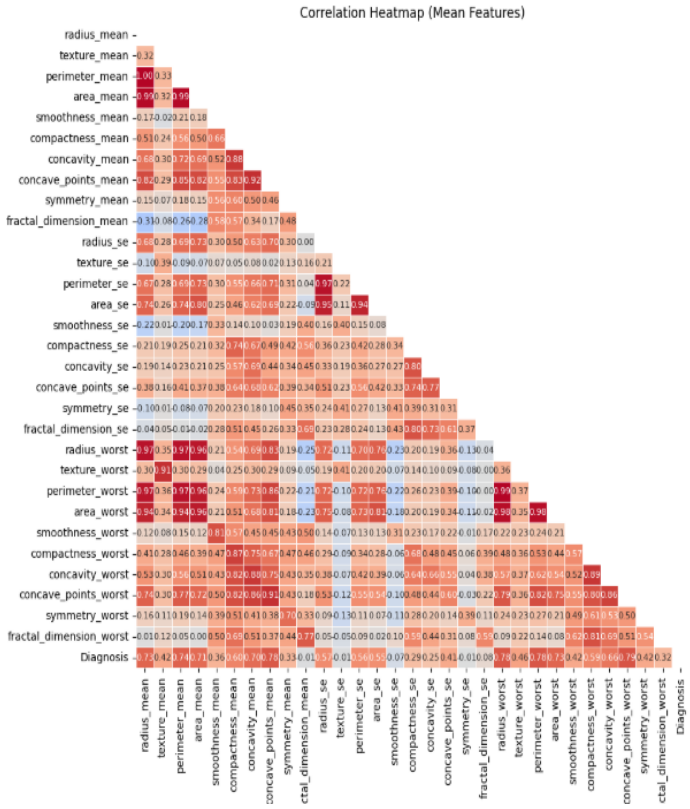

Fig. 1. Visualization of feature dependence

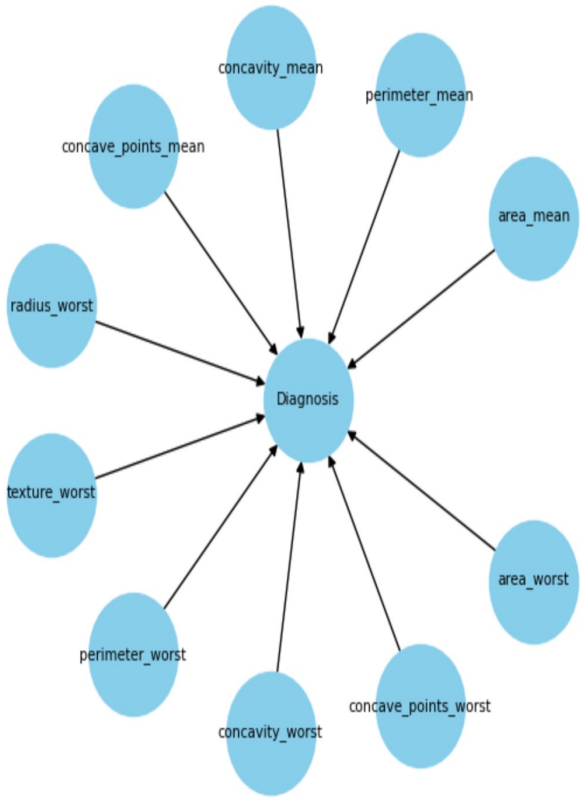

Fig. 2. Feature dependence

when the variable under examination is continuous, while a classification challenge occurs when the variable is categorical. Upon identifying the nature of the issue, the dataset gets splitted into training and testing sets, often with 20% or 30% test set. The model's performance is assessed using different assessment criteria which include acc, pr, re, and F1 score, to determine the best suitable deployment strategy.

D. Semi-supervised Learning

Semi-supervised learning improves the models efficiency by using the both labeled and unlabelled data.This type of technique is advantageous when their is a lack of labelled data . Initially the data is splitted and divided into 80 percent of training data and further the training data is splitted into two types labelled data and unlabelled data . Further the models are trained using labelled data and hyperparametrs are tuned to increase the efficiency.subsequently the model tries to guess the labels to be assigned to the unlabelled data,these predicted labels are again added to the labeled data. Finally again the hyperparameters are tuned to enhance the model efficiency.

E. Feature Extraction

1) Feature Agglomeration: Feature agglomeration is a technique within hierarchical clustering methods. This is an unsupervised feature extraction technique that effectively manages high-dimensional data to reduce complexity while preserving information [11].

Data points within a cluster should exhibit a high degree of similarity to each other, while remaining distinct from those in other clusters. This is the principle that regulates clustering. Hierarchical clustering is characterized by a tree-like structure and a decomposition methodology . It is essential to acknowledge that mistakes made at earlier stages cannot be corrected in later phases.

A bottom-up approach is employed in agglomerative clustering. This process operates under the assumption that the features are independent clusters. The clusters will subsequently merge based on proximity until the desired quantity of clusters is achieved. Subsequently, the clusters are characterized by features derived from pooled parameters, which, in this instance, will be ascertained using Ward's method, analogous to the mean. The transformation is achieved by replacing the original feature space with newly agglomerated features. Upon the fulfillment of all conditions, the process is eventually completed.

By consolidating related features into clusters, it enhances computational efficiency and result interpretability. The algorithm's hierarchical amalgamation of pairwise associations among features results in the decrease of overall quantity of composite features. This method aids in preserving the most significant areas of the original data through tasks such as grouping and classification.

2) Linear Discriminant Analysis (LDA): LDA establishes a new axis by maximizing the distance between class means and

by minimizing the variation among each class. It reduces data into a lower-dimensional space to improve class separation by generating linear discriminants that optimize the ratio for between-class and within-class variances. LDA efficiently handles cases where the no.of features are greater than the no.of samples and manages multicollinearity among features [12].

The feature extraction process involves finding a projection that maximizes the distance between class means represented by following objective function:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{1}$$

Where:

- w is the vector of weights representing the linear combination,
- S_b represents between-class scatter matrix,
- S_w represents within-class scatter matrix.

$$S_w = \sum_{i=1}^{k} S_i \tag{2}$$

$$S_i = \sum (x - \mu_i)(x - \mu_i)^T \tag{3}$$

$$S_b = \sum N_i (\mu - \mu_i)(\mu - \mu_i)^T \tag{4}$$

Where:

- N_i reprsents no.of data points in class i,
- μ_i represents avg vector of class i,
- μ represents total avg vector of all data points.

The aim is finding the vector W which maximizes the objective function, Which can be achieved by solving the following eigenvalue problem:

$$S_b W = \lambda_i S_w W \tag{5}$$

Where λ_i are the eigenvalues, and the corresponding eigenvectors are used to construct the transformation matrix. The top d eigenvectors (where d is the number of classes minus one) are selected to form the transformation matrix W:

$$W = [w_1, w_2, w_3, \dots] \tag{6}$$

Finally, the original data X is projected onto the new subspace:

$$X_{LDA} = XW \tag{7}$$

Where X_{LDA} is the extracted feature matrix.

F. Feature Selection

Feature selection is a major step in implementing the effective Machine learning models because it helps in decreasing the dimensionality and removing the unecessary features which may increase the complexity so in this study we have choose Bayesian Optimization and L2 regularization techniques for feature selection.

1) Bayesian optimization: Unlike other techniques which are compuatationally expensive Bayesian optimization works well in feature selction. Bayesian Optimization (BO) is a powerful technique which combines probabilistic modelling with an acquisition function where probabilistic modelling implemented as a Gaussian process (GP), this model delivers a posterior distribution over the function f based on observed data. Where the Acquisition Function describes the next following point for evaluation, with the expected improvement decision which is being most frequently utilized as it is optimally designed for performance enhancement in continuous domains with fewer than 20 dimensions and can consisits of stochastic noise in function evaluations. Bayesian optimization use probabilistic models to search, to facilitate adaptive sampling of hyper parameters and concentrate on attractive areas. Accordingly it decreases computing expenses, improves stability, and consists of the capability to yield an proper hyper-parameter configuration [13].

2) L2 regularization: Regularization is a technique employed to diminish model complexity by constraining the underlying functions, thereby enhancing generalization. Where one can ascertain the optimal regularization by evaluating the trade-off between variance and bias. The experimental results indicate that the L2 parameter regularization method, commonly referred to as Ridge regularization, is the most effective. Ridge regularization employs an L2 penalty based on the squared magnitude of the coefficients. This approach mitigates the risk of overfitting, identifies critical components, and improves model stability. This method effectively diminishes the variance of the coefficients linked to the regression model by imposing restrictions on them. L2 regularization, unlike Lasso regularization, does not promote sparsity; instead, it normalizes the coefficients towards zero without rendering them null. Lasso regularization facilitates the promotion of sparsity. L2 regularization helps in balancing between the underfitting and overfitting in ML and also enhancing overall performance. it tries to generalizes and understands the patterns in the data by overcoming the problem of byhearting. Furthermore, regularization can enhance model stability by diminishing the coefficients' susceptibility to minor data perturbations. This technique is especially beneficial in scenarios where features demonstrate significant correlation, referred to as multicollinearity.

G. Classification Models

In this research work, a range of different classification algorithms were utilized to predict the outcome of breast cancer diagnosis [14], comparing both supervised learning and Semi-Supervised Learning approaches. The models employed include LR, Gaussian-NB, Linear SVM, RBF SVM, Decision tree, Random forest, X-GB, Gradient Boost& K-Nearest Neighbors. Supervised learning models rely on fully labeled training data to learn a classification function, while semi-supervised learning incorporates both labeled & unlabeled data, leveraging the structure within the data to improve model performance, especially when labeled data is limited [15].

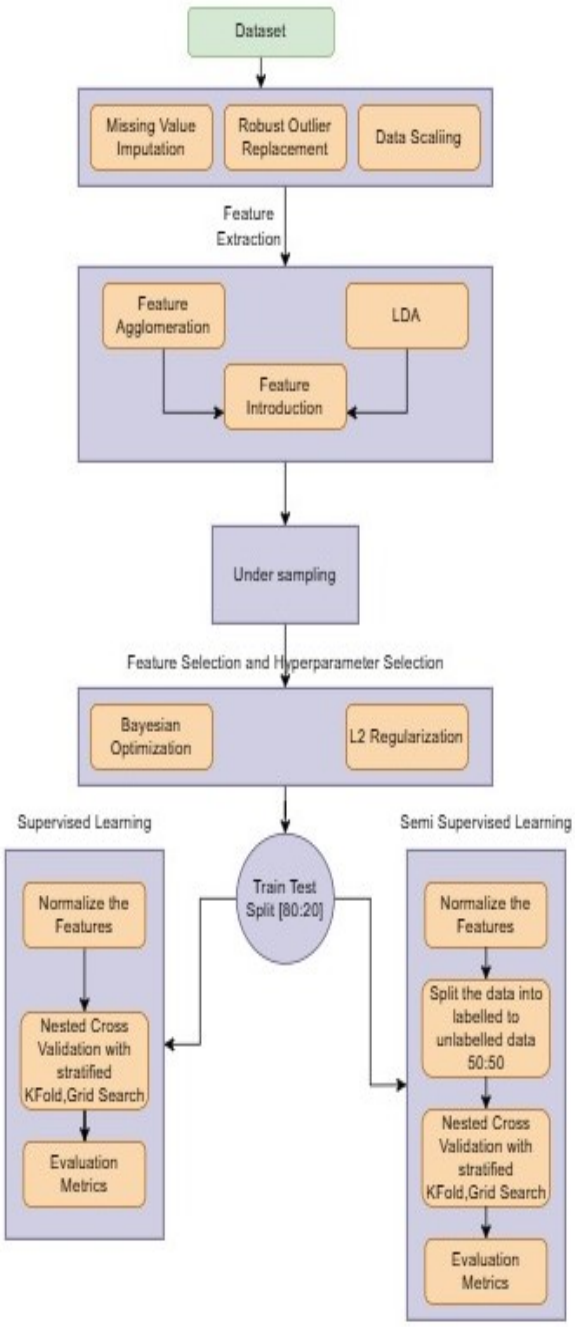

Fig. 3. WorkFlow

Notably, semi-supervised models achieved comparable or even superior performance over their supervised counterparts [16], highlighting the effectiveness of utilizing unlabeled data in enhancing classification accuracy [17] [18].

H. Proposed Approach

We began the setting up of a feature extraction strategy through horizontal stacking utilizing Factor Analysis (FA) and Linear Discriminant Analysis (LDA) methodologies. Both FA and LDA methodologies were independently employed on the preprocessed dataset to extract significant features and minimize data dimensions. Then features drawn from each of the methods were stacked horizontally, that gave us an aggregated dataset with reduced dimensions. This method helped us to use the individual advantages of both FA and LDA in feature extraction Subsequently thus feature selection and regularization methods were utilized to enhance the performance. Bayesian optimization was employed to identify the most significant and useful features through a systematic exploration of the feature space. Additionally, L2 regularization was implemented to penalize less significant features thereby fostering a more robust and streamlined model. This step highlighted the critical information in the data and enabled the identification of the hyperparameters which are necessary for model training. At-last GridSearch was applied to optimize the model's performance through the systematic evaluation of various hyperparameter combinations. This optimization process enabled us to get the optimal parameters for each model,which ensures us that the output was maximally accurate and efficient.

IV. EVALUATION ANALYSIS

A. Evaluation Metrics

Classification algorithm's performance is usually evaluated using different metrics such as acc, pr, re & F1 score. To evaluate the model performance dataset is divided into training set and testing set. The metrics are tested on the testing dataset. The confusion matrix includes True-positive, True-negative, False-positive, and False-negative which are used to compute evaluation metrics.

$$\text{Accuracy} = \frac{T_p + T_n}{T_n + T_p + F_n + F_p} \tag{8}$$

$$\text{Sensitivity} = \frac{T_p}{T_p + F_n} \tag{9}$$

$$\text{F1 score} = 2\left(\frac{P \cdot R}{P + R}\right) \tag{10}$$

$$\text{Specificity} = \frac{T_p}{T_p + F_p} \tag{11}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \tag{12}$$

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{13}$$

where Tp = instances which are correctly identified as positive, Fp = instances which are wrongly identified as positive, Fn = instances which are wrongly marked as negative, Tn = instances which are correctly marked as negative.

B. Experimental Results

In this study, we conducted experiments on various models with the primary goal of identifying a workflow that yields an accurate breast cancer diagnosis using SL and SSL methodology. when we look into the results on WBCD We got an average accuracy of 94.14 % for SL and 94.50% for SSL where Random Forest achieves the highest accuracy (95.70%), indicating its robustness in semi-supervised settings.Both Logistic Regression and Linear SVM perform consistently well across all learning types, 94.62%. Gaussian Naive Bayes performs the worst, at 92.47% for both SL and SSL, suggesting it is not particularly suited to this data. Overall, the SSL models performed slightly better on prediction.

When it comes to (WDBC) got an average accuracy of 98.83 % for SL and 98.73 for SSL and most of the model achieved 99.12 in both methodologies but Gaussian Naive Bayes and KNN not performed well in SL

The experiments demonstrated that with the proper combination of preprocessing techniques and machine learning methods SSL can be applied as efficient option for accurate breast cancer diagnosis rather than SL.

C. Discussion

The primary objective of this study was to find alternative approach for supervised learning that is semi-supervised learning (SSL) for diagnosis of breast cancer. This is advantageous in situations when labelled data is limited or labelling is resource-intensive. Considering that real-world data frequently lacks labels, manual labelling is laborious, requiring substantial time and human resources.

Our results show that SSL, when combined with the proposed methodology, can produce results that are comparable to or even better than those obtained using SL algorithms. This implies that SSL offers a efficient way of automating the diagnostic process, especially in medical applications where data labeling is a major bottleneck. The comparative performance of SL and SSL algorithms is summarized in the table below:

V. Conclusion

From this study we have concluded that accuracies of Semi-supervised learning and supervised learning are mostly near. All the accuracies of the algorithms with these techniques are almost in the range of 91 to 99 percentage. From this study, we also ensure that the models didn't undergo underfitting nor overfitting, providing the accurate predictions for both the BENIGN and Malignant tumors. This is shown in the figures. Semi-supervised learning provides us with an advantage over supervised learning when the labeled data is scarce. Recent studies, such as Wang et al., have showcased that SSL can leverage large amounts of unlabelled data with low computational requirements compared to SL. This makes SSL a better alternative for SL when labeled data is expensive or requires more computational resources. Since we have achieved 98% accuracy, there can be future advancements in improving the

Model	Acc	Pr	Re	F1
LR	99.12%	99.32%	98.81%	99.05%
GNB	98.25%	98.65%	97.62%	98.10%
SVM-L	99.12%	99.32%	98.81%	99.05%
SVM-RBF	98.25%	98.65%	97.62%	98.10%
DT	99.12%	99.32%	98.81%	99.05%
RF	99.12%	99.32%	98.81%	99.05%
XGB	99.12%	99.32%	98.81%	99.05%
GB	99.12%	99.32%	98.81%	99.05%
KNN	98.25%	98.65%	97.62%	98.10%

TABLE I
SL Models Metrics for WISCONSIN(diagnostic)

Model	Acc	Pr	Re	F1
LR	99.12%	99.32%	98.81%	99.05%
GNB	97.37%	98.00%	96.43%	97.13%
SVM-L	99.12%	99.32%	98.81%	99.05%
SVM-RBF	99.12%	99.32%	98.81%	99.05%
DT	98.25%	98.12%	98.12%	98.12%
RF	99.12%	99.32%	98.81%	99.05%
XGB	99.12%	99.32%	98.81%	99.05%
GB	99.12%	99.32%	98.81%	99.05%
KNN	98.25%	98.65%	97.62%	98.10%

TABLE II
SSL Models Metrics for WISCONSIN(diagnostic)

Model	Acc	Pr	Re	F1
LR	94.62%	94.62%	94.58%	94.61%
GNB	92.47%	92.47%	92.43%	92.46%
SVM-L	94.62%	94.62%	94.58%	94.61%
SVM-RBF	93.55%	93.67%	93.47%	93.53%
DT	93.55%	94.62%	93.61%	93.55%
RF	94.62%	95.70%	94.58%	94.61%
XGB	94.62%	94.62%	94.58%	94.61%
GB	94.62%	94.62%	94.58%	94.61%
KNN	94.62%	94.62%	94.58%	94.61%

TABLE III
SL Models Metrics for WISCONSIN(original)

Model	Acc	Pr	Re	F1
LR	94.62%	94.62%	94.58%	94.61%
GNB	92.47%	92.47%	92.43%	92.46%
SVM-L	94.62%	94.62%	94.58%	94.61%
SVM-RBF	94.62%	94.62%	94.58%	94.61%
DT	94.62%	94.62%	94.58%	94.61%
RF	95.70%	95.70%	95.69%	95.69%
XGB	94.62%	94.62%	94.58%	94.61%
GB	94.62%	94.62%	94.58%	94.61%
KNN	94.62%	94.62%	94.58%	94.61%

TABLE IV
SSL Models Metrics for WISCONSIN(original)

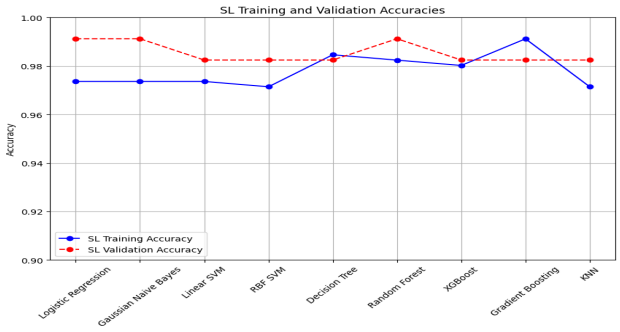

Fig. 4. SL Training and Validation Accuracies

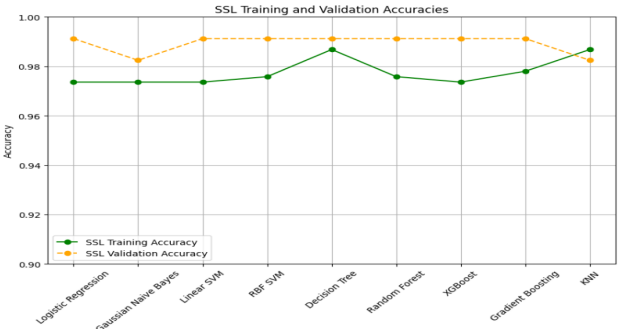

Fig. 5. SSL Training and Validation Accuracies

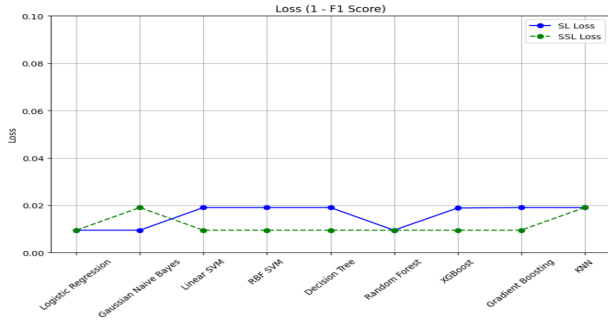

Fig. 6. Loss for SL and SSL models

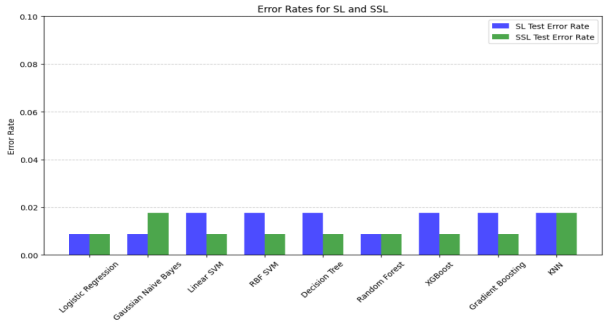

Fig. 7. Error Rates for SL and SSL models

accuracy of the model by using different techniques like cross-validation and hyper-parameter tuning.

REFERENCES

[1] Amreen Batool and Yung-Cheol Byun. Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm. *IEEE Access*, 12:12869–12882, 2024.

[2] Mohammad Reza Darbandi, Mahsa Darbandi, Sara Darbandi, Igor Bado, Mohammad Hadizadeh, and Hamid Reza Khorram Khorshid. Artificial intelligence breakthroughs in pioneering early diagnosis and precision treatment of breast cancer: A multimethod study. *European Journal of Cancer*, 209:114227, 2024.

[3] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, 13(17):4287, 2021.

[4] Prerna Singh, Bhumika Minhas, Sejal, and Pulkit Dwivedi. A semi-supervised quantitative inference model for accurate breast cancer detection. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, volume 6, pages 604–609, 2023.

[5] Mostafa Shanbehzadeh, Hadi Kazemi-Arpanahi, Mohammad Bolbolian Ghalibaf, and Azam Orooji. Performance evaluation of machine learning for breast cancer diagnosis: A case study. *Informatics in Medicine Unlocked*, 31:101009, 2022.

[6] K. Rustagi, P. Bhatnagar, R. Mathur, I. Singh, and S. K. G. Hybrid salp swarm and grey wolf optimizer algorithm based ensemble approach for breast cancer diagnosis. *Multimedia Tools and Applications*, 83(27):70117–70141, 2024.

[7] S. Prusty, P. Das, S. K. Dash, S. Patnaik, and S. G. P. Prusty. Retracted: Prediction of breast cancer using integrated machine learning-fuzzy and dimension reduction techniques. *Journal of Intelligent & Fuzzy Systems*, 45(1):1633–1652, 2023.

[8] K. Parashar, A. Kaushik, R. Sharma, and N. Aman. Breast tumor prediction using svm with rain fall optimisation algorithm. In *Smart Innovation, Systems and Technologies*, pages 167–180, 2024.

[9] V. Nemade, S. Pathak, and A. K. Dubey. Hybrid deep convolutional neural network approach for detecting breast cancer in mammography images. *International Journal of Electrical and Electronics Engineering*, 10(5):102–119, 2023.

[10] UCI Machine Learning Repository. Breast cancer wisconsin (diagnostic) data set, n.d. Accessed: 2024-01-25.

[11] Ankit Jalan and Purushottam Kar. Accelerating extreme classification via adaptive feature agglomeration. pages 2600–2606, 08 2019.

[12] Elhadji Ille Gado Nassara, Edith Grall-Maës, and Malika Kharouf. Linear discriminant analysis for large-scale data: Application on text and image data. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 961–964, 2016.

[13] Achala Shakya, Mantosh Biswas, and Mahesh Pal. Chapter 9 - classification of radar data using bayesian optimized two-dimensional convolutional neural network. In Prashant K. Srivastava, Dileep Kumar Gupta, Tanvir Islam, Dawei Han, and Rajendra Prasad, editors, *Radar Remote Sensing*, Earth Observation, pages 175–186. Elsevier, 2022.

[14] Sreedhar Kollem, Chandrasekhar Sirigiri, and Samineni Peddakrishna. A novel hybrid deep cnn model for breast cancer classification using lipschitz-based image augmentation and recursive feature elimination. *Biomedical Signal Processing and Control*, 95:106406, 2024.

[15] Fatima-Zahrae Nakach, Ali Idri, and Evgin Goceri. A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. *Artificial Intelligence Review*, 57(12):327, 2024.

[16] Md Rakibul Islam, Md Mahbubur Rahman, Md Shahin Ali, Abdullah Al Nomaan Nafi, Md Shahariar Alam, Tapan Kumar Godder, Md Sipon Miah, and Md Khairul Islam. Enhancing breast cancer segmentation and classification: An ensemble deep convolutional neural network and u-net approach on ultrasound images. *Machine Learning with Applications*, 16:100555, 2024.

[17] Oumeima Thaalbi and Moulay A. Akhloufi. Deep learning for breast cancer diagnosis from histopathological images: classification and gene expression: review. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1):52, 2024.

[18] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, and J. Zhang. Deep learning models for breast cancer detection and diagnosis: A review. *Medical Image Analysis*, 80:102064, 2024.