Importance of Data Scaling for Predicting Diabetes and Breast Cancer with Naïve Bayes and Back Propagation Neural Network

Issac. P. J[1], Dr. G R Sridhar[2]

[1]Research Scholar, Computer Science, Rayalaseema University, Kurnool, AP
[2]Co-Author, M.B.B.S., MD, DM, Director, Endocrine and Diabetes Centre, Visakhapatnam, AP 530002

Abstract:
Since predicting chronic diseases is essential, a number of methods are being developed. Diabetes and Cancer are some of the most common non-communicable diseases causing disability and death. Hence early detection methods are invaluable for these two. In this research, three different classification techniques are implemented: Back Propagation Neural Network (BPNN), Naïve Bias (NB), and a combination of both algorithms to predict breast cancer and diabetes. The proposed technique is evaluated for different metrics like accuracy, precision, recall, and false positive rate with both scaled data and the original dataset. For diabetes, an accuracy of 88% was obtained in the actual dataset, whereas the scaled dataset has obtained an accuracy of 93.72%. Similarly, for the breast cancer dataset, an accuracy of 91% was obtained in the actual dataset, whereas the scaled dataset has obtained an accuracy of 93.57%. It is concluded that the scaled data provides better performance metrics. Also, the ensemble approach provides better performance metrics than the individual models for both diabetes and breast cancer.

Keywords: Disease Detection, Disease Prediction, Machine Learning, Neural Network, AI, Back Propagation Neural Network (BPNN), Naïve Bias (NB), Ensemble Algorithm, Breast Cancer, Diabetes

1 Background

The leading chronic diseases are diabetes and cancer. Early diagnosis of these chronic diseases allows early treatment and prevention. (Brady et al., 2021). Clinical databases contain disease-related traits that may be anticipated before the diseases actually manifest and enable better treatment at early stage.

Chronic illnesses develop over a lengthy period of time, and can advance slowly and vary in severity from person to person (Cambridge university Press, 2014). As a non-communicable illness, it is a primary cause of shorter lifespan globally. Multi-morbidity rates based on chronic conditions rose with age in industrialised nations (Kuipers et al., 2020). Multi-morbidity affects around 40% of the overall world population. Chronic illness management is a difficult issue in any healthcare system because it requires both acute episodic treatment and long-term care planning (Corbett et al., 2020). Chronic illnesses need long-term management, care, and surveillance. For efficient chronic illness management, healthcare systems must have coordination, comprehensiveness, and continuity. The necessity of health policy is also stressed, requiring proactive rather than reactive health management. Those nations with primary care require a low-cost health monitoring system (McGetrick et al., 2019; Sporinova et al., 2019).

Artificial Intelligence can be used to make early prediction by using the data from electronic medical records of the patients. Electronic Health Records (EHRs) are generally, digital versions of paper-based medical records, where the patient's medical history is included, from daily physical tests and diagnosis through treatments (Anshari, 2019). It is common for hospitals and other healthcare organisations to gather and keep this data. They are more valuable than paper records because they let physicians and other medical professionals keep track of patients over time, choose patients for follow-up visits and treatment options, and evaluate patients. The advantages of electronic medical records are well established. With the use of all this patient data, it may be possible to predict diseases in new patients (Hang et al., 2019). Machine learning is becoming increasingly widespread in the healthcare profession as the quantity of data grows. Data from electronic health records (EHRs) may be analysed using machine learning (ML) techniques to uncover hidden patterns in the data and predict outcomes (Jain, 2019).

Medical records have become very difficult to handle due to a number of technological advancements in data processing. Clustering Algorithm provides centroid-based clustering efficiency for a limited number of Fuzzy C-means samplings (Li et al., 2019). In the midst of health emergencies, machine learning algorithms can help providing speedy and accurate diagnosis (Chen et al., 2019).. Besides, incorporating machine learning-based modelling for administrative data sets may assist predict possible problems, minimise the consumption of health care resources and personalise outcomes.

A medical database may have duplicate data and missing values, making data mining difficult. This impacts the mining functionality and must be fixed for proper data preparation and reduction by data mining algorithms. Data must be exact, consistent, and devoid of noise to forecast illnesses. Chronic Disease Diagnosis (CDD) is a useful method for managing chronic illnesses. This helps doctors and patients offer healthcare efficiently (Jain & Singh, 2018)

Tseng et al. (2019) has investigated the epidermal variables related in breast cancer metastatic prediction. This study employs support vector machines, logistic regression, Bayesian classification, and random forest classifiers to analyse data. A total of 302 cancer patients were used for the study. The results have shown that random forest classifier accurately predicts breast cancer in the recent 3 months. This study revealed that early detection and adequate treatment of cancer are required for successful functional characteristics performance. This study focused on cancer diagnosis but neglected categorization.

Singh (2019) has focused on clinical characteristics with regular blood samples. This study uses anthropometric measures to assess breast cancer. This study also uses machine learning to assess the role of machine learning in cancer prediction by considering feature selection, classification, and data division protocols. This study evaluated clinical databases and anthropometric measurements from breast cancer patients. This research evaluated several attributes of clinical database through consideration of several factors. The statistical factors and other qualities are used to assess the specific features' relevance. The classification methods used in this study include logistic regression, linear discriminant, random forest, support vector machine (SVM), quadratic discriminant, and K-nearest neighbour (K-NN). For example, age as a biomarker for breast cancer is one of nine variables assessed for disease prediction. Furthermore, K-NN provides 92.105 % classification accuracy while medium Gaussian SVM provides 83.684 %. No mathematical derivation of the classifier approach is offered in this study.

LaPierre et al. (2019) has reviewed the methods and algorithms adopted for the deep learning process.

For analysis this research considered Type 2 diabetics dataset for processing and evaluated for the deep learning process. Results showed that machine learning algorithm performs effectively for extraction of features and learning characteristics. Through analysis this research concluded that machine learning performs effectively for the clinical database in disease prediction. However this research subjected serious limitation of it does not offer any specific information about machine technique which performs effectively.

Zhang et al. (2019) has proposed two different weighting criteria which operate based on the consideration of area under the ROC curve (AUC) and the concordance index (C-index), respectively. Also this research uses Dirichlet-based regularization for weights calculation to estimate the difference between the predictive ability of the model. Experimental results stated that Centre Hospitalier Universitaire de Sherbrooke (CHUS) exhibits an effective machine learning approach for medical application. But this research has failed to provide a detailed explanation about the classification approach.

Medical diagnosis utilising ML approaches has acquired great traction in the last decade. This development in the application of ML approaches is partially attributable to the fact that it gives better diagnosis of numerous illnesses, ascribed to greater symptom detection. Further, machine learning also helps to create a more tailored therapy proposal, as findings of the analysis may be utilised to present new diagnostic hypotheses (Houfani et al., 2020). Breast cancer screening approaches have been applied employing ML algorithms. The most generally used and the most beneficial ones are neural networks and Decision trees (Kaushal & Singla, 2021). However, their operating methods are completely different. The key advantage of utilising a decision tree-based ML model is that the technique utilised is straightforward to grasp and has also been demonstrated to be highly efficient.

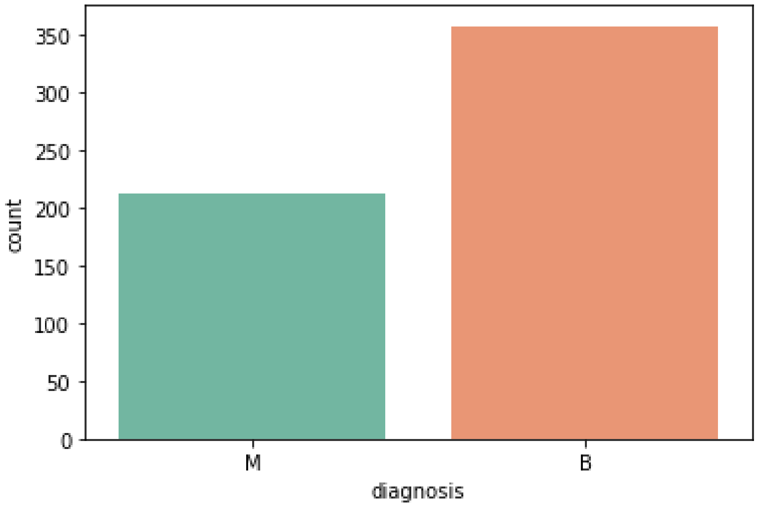
2 Methodology

The chronic diseases can be predicted at early stages by using the proposed model described in this section. While any chronic disease can be identified based on the available of dataset, this research will focus on predicting early stage breast cancer and diabetes. The data used for this research involves electronic medical records which will contain various features that may or may not be dependent on the disease. The dataset used for diabetes is obtained from Kaggle (Singh, 2017), while the dataset for Breast cancer is taken from Breast Cancer Wisconsin (Diagnostic) Data Set (Learning, 2016) The relevant features must be identified by the model and consider it for the classification. The final classification utilises both BPNN and Naïve Bayes to identify the disease, and to obtain the performance metrics. These algorithms are used in parallel, and also as an ensemble approach.

Since there are no datasets which contains data for both cancer and diabetes, two different datasets are considered for the review. The Electronic Medical Records (EMR) is collected from various health centres and hospitals, and compiled together according to the common variables. The dataset is split into training and testing datasets. To train and test the dataset, a certain ratio (70:30) of the overall dataset is used. There's no need to use test data during training. Also, more data are utilised for training than for testing.

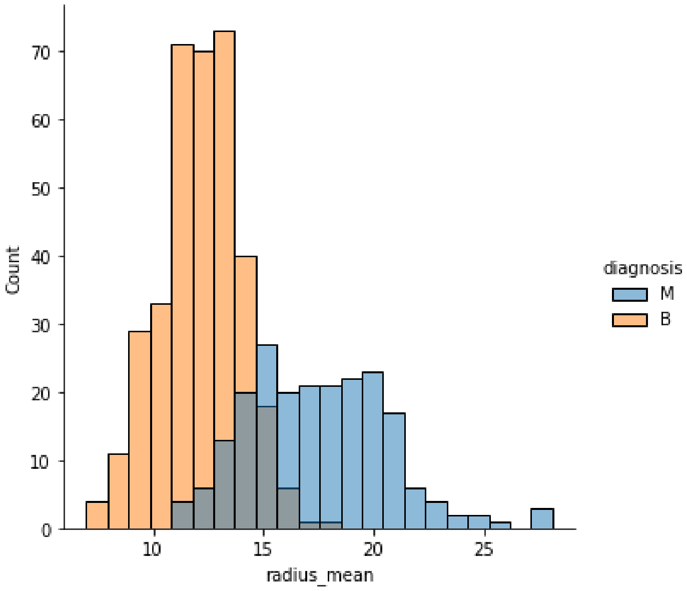
It is essential to comprehend the data to check the necessary features in the dataset. In the case of Breast cancer dataset Breast Cancer Wisconsin (Diagnostic) Data Set (Learning, 2016), the target variable “Diagnosis” is first analyzed, and the number of features are counted to check the dataset is balanced. In the data, there are totally 569 rows of data, out of which 357 rows are marked as benign (B), while 212 instances are marked as malignant (M) as shown in figure 1.

Figure 1 Target Variable



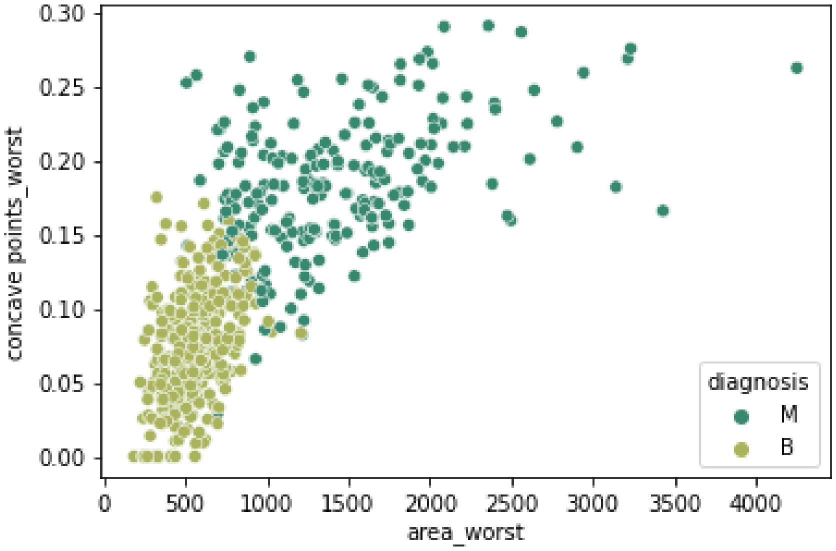
Therefore, the number of patients with breast cancer are lesser than those without breast cancer; this might lead into a misclassification. However, since the difference between the counts is not very much we can proceed further. After looking at the features in the dataset, another important variable that may affect the outcome of the results is the mean radius value. It is the average distance between the centre and the perimeter of the breast. The density of this variable is shown in figure 2

Figure 2 Density of Radius Mean feature



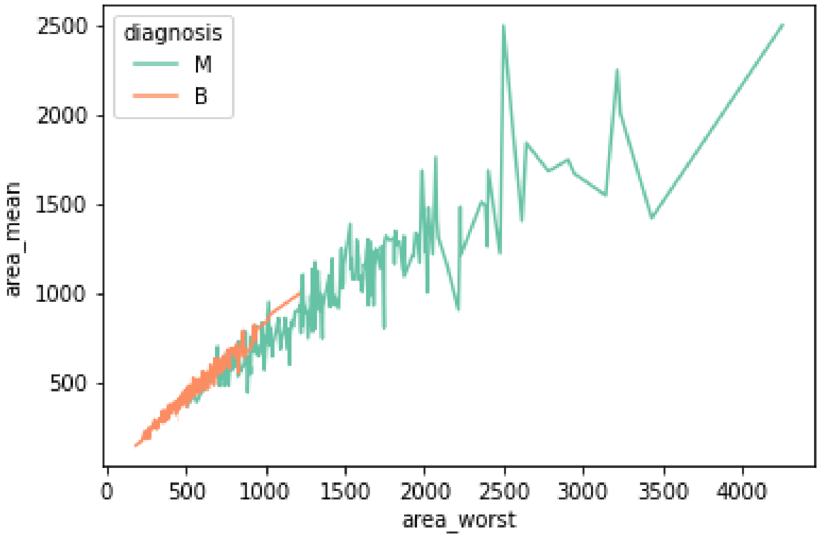
The important features that have higher correlation are now looked between each other for cross correlation. Figure 3 displays the cross correlation between area_worst and concave points_worst variables with respect to the target variable “Diagnosis”.

Figure 3 Cross Correlation between area_worst and concave points_worst



From the plot, the separation between the diagnoses is clearly visible. It is inferred that these two features are important in classifying the patients whether they have breast cancer or not. However, since they are completely correlated, one of these features may have to be deleted since it might not be useful to have both of them. However, both these features are important while performing the feature selection, hence both are used for the next stage. Similarly, cross correlation is performed between concave area_mean and area_worst with respect to the target variable Diagnosis as shown in figure 4. This plot classifies the diagnosis feature but not as much as the previous plot. The points in the plot are more spread out, where both the features are in different dimensions.

Figure 4 Cross Correlation between area_worst and area_mean



In this research, a 70:30 split between training and testing data is established. From the diabetes dataset, Kaggle (Singh, 2017), there are 788 rows of data out of which 552 rows are used for training the classifiers, while 236 rows of data are used for the actual testing and evaluation. Similarly, from the breast cancer dataset, there are 569 rows of data out of which 399 rows are used for training the classifiers, while 171 rows of data are used for the actual testing and evaluation. Some of the features in cancer dataset are Compactness, Concavity, and Concave points, while some of the features from diabetes dataset are Skin Thickness, Insulin, and BMI.

Table 1 Training and Testing Data Size

	Cancer	*Diabetes*
Total Size	569	788
Training	398	552
Testing	171	236

Using a variety of methods, the data is pre-processed. Before being categorised, the text in the dataset must be cleaned and processed. To start with, any unnecessary data is cleaned up by removing any background noise from the document. Linear transform, a statistical transform, is used to reduce the noise. Because of this, the classifier will place a greater emphasis on the distribution of data points. Because there will be a large number of irrelevant variables, only those that are essential to the prediction process need to be extracted. An approach known as Principle Component Analysis (PCA) is used in this study in order to locate and extract the most important characteristics. Since certain features may contain more data than others, the data is first scaled. Consequently, the features with a smaller amount of data will be scaled in order to ensure that all features are represented equally. By employing normalisation, unnecessary data can be removed and replaced with the more relevant information.

Scaling is used as a means of standardisation. Under some circumstances, scaling is very advantageous and necessary, particularly when the data parameters cover a wide range (Ahsan et al., 2021). There are a variety of ways in which scaling may be implemented. This research employs a range scaling type that shifts and amplifies the values into a new range. To identify the best features, the scaled data are input into PCA after scaling. Classification is the next stage in the process of advancement.

For more accurate prediction, the retrieved data must be categorized. In order to classify the data, the pre-processing procedure must be finished first BPNN and Naive Bayes are thus combined in a hybrid approach. In addition, a gradient boosting approach is used to enhance the quality of each iteration. Using the dataset, we train the two optimization approaches together. The pre-processed and feature-extracted data is supplied into the classifiers for accurate prediction after the training procedure.

3 Results and Discussion

The algorithm is implemented using python programming and the following results are obtained. The dataset is split into training and testing dataset with training dataset being larger than the testing dataset at a 70:30 ratio. Out of the two algorithms Naïve Bayes, and Back Propagation Neural Network implemented for Diabetes data, it is seen that BPNN with 84% has a comparatively higher accuracy than Naïve Bayes algorithm with 83%as shown in table 2. Similarly, the precision, recall and F-measure are also higher for BPNN algorithm. When the algorithms are combined into an ensemble model, the overall accuracy is further increased to 88%.

Table 2 Evaluation without Scaled Data- Diabetes

Diabetes Prediction (Without Scaling the Data)

Evaluation Metrics	***NB***	***BPNN***	***Ensemble***
Accuracy	0.83	0.84	**0.88**
Precision	0.77	0.79	**0.85**

Recall	0.72	0.75	**0.79**
F-Measure	0.75	0.77	**0.82**

Data scaling is now performed on the dataset and the model is run for both the algorithms, and for the ensemble algorithm. It is seen that BPNN with 92% has a comparatively higher accuracy than Naïve Bayes algorithm with 90%as shown in table 3. Similarly, the precision, recall and F-measure are also higher for BPNN algorithm. When the algorithms are combines into an ensemble model, the overall accuracy is further increased to 94%. Also, it can be seen that the scaled data has significantly higher performance metrics than the one without scaling. This shows that data scaling significantly improves the performance of the classifiers for disease prediction

Table 3 Evaluation with Scaled Data- Diabetes

Diabetes Prediction (Scaling the Data)

Evaluation Metrics	*NB*	*BPNN*	*Ensemble*
Accuracy	0.90	0.92	0.94
Precision	0.86	0.88	0.89
Recall	0.86	0.89	0.93
F-Measure	0.86	0.88	0.91

Similar implementation is performed now for the cancer dataset. Out of the two algorithms Naïve Bayes, and Back Propagation Neural Network implemented for cancer data, it is seen that BPNN with 87% has a comparatively higher accuracy than Naïve Bayes algorithm with 84%as shown in table 4. Similarly, the precision, recall and F-measure are also higher for BPNN algorithm. When the algorithms are combines into an ensemble model, the overall accuracy is further increased to 91%.

Table 4 Evaluation without Scaled Data- Breast Cancer

Breast Cancer Prediction (Without Scaling the Data)

Evaluation Metrics	*NB*	*BPNN*	*Ensemble*
Accuracy	0.84	0.87	0.91
Precision	0.88	0.91	0.93
Recall	0.87	0.89	0.93
F-Measure	0.87	0.90	0.91

Data scaling is now performed on the Cancer dataset and the model is run for both the algorithms, and for the ensemble algorithm. It is seen that BPNN with 92% has a comparatively higher accuracy than Naïve Bayes algorithm with 91%as shown in table 5. Similarly, the precision, recall and F-measure are also higher for BPNN algorithm. When the algorithms are combines into an ensemble model, the overall accuracy is further increased to 94%. Also, it can be seen that the scaled data has significantly higher performance metrics than the one without scaling. This shows that data scaling significantly improves the performance of the classifiers for disease prediction

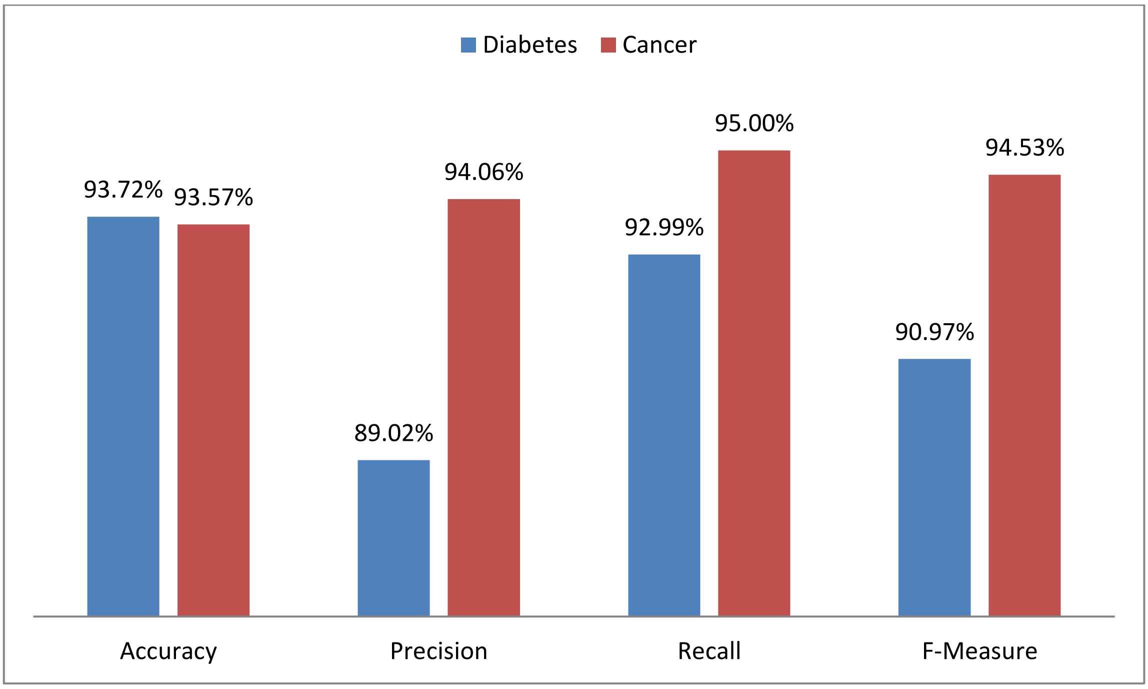
Table 5: Evaluation with Scaled Data- Breast Cancer

Breast Cancer Prediction (Scaling the Data)

Evaluation Metrics	*NB*	*BPNN*	*Ensemble*
Accuracy	0.91	0.92	0.94
Precision	0.92	0.93	0.94
Recall	0.93	0.94	0.95
F-Measure	0.93	0.93	0.95

The ensemble algorithm for the scaled data has comparatively performance metrics for both the datasets. The classifiers were compared "without scaling" and "with scaling" for both types of diseases. The evaluation parameters/factors of the three classifiers were compared with and without data scaling. Finally, the suggested classifier's overall improvement over the previous two classifiers was evaluated. The naive Bayes prediction model has 192 accurate classifications and 39 misclassifications without scaling the data. The scaled data included 44 incorrect predictions/classifications. Scaled data increased this model's accuracy by 8.85 %, precision by 11.96 %, recall by 19.54 %, and F-measure by 15.75 %. Using scaled data lowered the error rate by 39.52 %, the specificity by 4.39 %, and the false positive rate by 34.88 %. When no scaling was used for diabetes prediction, there were 36 misclassifications and 195 accurate predictions in the BPNN model. Scaling revealed 38 incorrect predictions and 424 correct classifications with a test data size of 462. The BPNN prediction model's accuracy improved by 8.71 %, precision by 11.326 %, recall by 18.747 %, and F-measure by 15.02 %. Also, the error rate was dropped by 37.24 % , the specificity was enhanced by 4.39 % , and the false positive rate was reduced by 37.32 % by employing scaled data.

Figure 5 Comparison of accuracy for scaled data.



The ensemble prediction model generated 203 correct classifications and 28 incorrect classifications without scaling the data. The scaled data revealed 433 incorrect predictions/misclassifications and 29 false predictions/misclassifications. With scaled data, this model's accuracy, precision, recall, and F-measure increased by 8.85, 11.96, 19.54, and 15.75 %. Using scaled data lowered the error rate by 39.70%, increased specificity by 1.49%, and decreased false positives by 18.96%. The effectiveness of the prediction model was determined by the number of incorrect positive and negative predictions produced. The suggested ensemble classifier outperformed NB and BPNN classifiers with and without data scaling in terms of accuracy and performance. It improved DIABETES prediction by 4.09 to 10.53 % without scaling data and 1.65 to 7.82 % with scaled data.

Similarly, for breast cancer, the naive Bayes model produced 144 accurate classifications and 27 incorrect predictions without scaling the data. In the case of non-scaled data, there were 171 predictions. With scaling, 342 predictions were produced, 312 of which were right. In addition, scaling lowered the mistake rate and false positive rate. Improvements in accuracy, precision, recall and F-measure were all 8.34% higher than before. The error rate was lowered by 44.46 %, specificity increased by 11.01 %, and false positives were reduced by 42.37 %. For a total of 171 predictions, the BPNN model had 149 right classifications and 22 wrong predictions. Total predictions modified to 342 with 315 right classifications and 27 incorrect classifications with a scaling rate of 2. Scaling helps improve accuracy, precision, recall, and F-measure values. These indexes improved by 5.72, 2.21, 5.71, and 3.96 %, respectively. The following parameters lowered the mistake rate and false positive rate: 38.64% and 33.92% with 6.40% specificity.

For a total of 171 predictions, the ensemble model showed 156 correct predictions and 15 incorrect classifications. The total number of predictions was 342 with a scaling factor of two. Assume the ensemble model makes 320 correct predictions but 22 incorrect ones. Including in the previous situations, scaling all parameters across a similar range aided typical performance evaluation metrics like accuracy, precision, recall, and F-measure augmentation. The %age improvements are 2.56, 0.64, 2.60, and 1.62. Also, the error rate and false positive rates were lowered. Both specificity and reduction %ages improved by 2.99 %. Each false positive or negative affects the classification performance. The model's efficiency decreases with the number of positive and negative false alarms. The ensemble model was presented to reduce FP and FN. The suggested ensemble model outperformed the NB and BPNN models in terms of accuracy, precision, recall, and error rates. It improved BC prediction by 3.19 to 8.33 % when data were scaled, and by 1.09 to 2.65 % when data were unscaled. We observed that scaling improved performance more in Naive Bayes than BPNN, but less in ensemble than the other two models. Because the ensemble model had good performance with few misclassifications, scaling only modestly improved it further.

4 Conclusion

In this research three models have been implemented individually, which are Naïve Bayes algorithm, BPNN algorithm, and the combined Ensemble algorithm. The data is analysed using these three models and various performance metrics like accuracy, prediction, recall, and F-measure have been obtained. A comparative analysis with each classifier was carried out for both diabetes and cancer, and also analysed with scaled data and without scaling. It is seen from the results that data scaling significantly improved the accuracy. The ensemble classifier has better classification results when compared to the individual algorithms for both diabetes and breast cancer. For diabetes, an accuracy of 88% was obtained in the actual dataset whereas the scaled dataset has obtained an accuracy of 93.72%. Similarly, for breast cancer dataset, an accuracy of 91% was obtained in the actual dataset whereas the scaled dataset has obtained an accuracy of 93.57%.

References

1. Ahsan, M.M., Mahmud, M.A.P., Saha, P.K., Gupta, K.D. & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9 (3). pp. 52. DOI: 10.3390/technologies9030052.
2. Anshari, M. (2019). Redefining Electronic Health Records (EHR) and Electronic Medical Records (EMR) to Promote Patient Empowerment. IJID (International Journal on Informatics for Development), 8 (1). pp. 35. DOI: 10.14421/ijid.2019.08106.
3. Brady, A.M., Deighton, J. & Stansfeld, S. (2021). Chronic illness in childhood and early adolescence: A longitudinal exploration of co-occurring mental illness. Development and Psychopathology, 33 (3). pp. 885–898. DOI: 10.1017/S0954579420000206.
4. Cambridge university Press (2014). Coping with chronic illness from Psychology, health and illness. 2014.
5. Chen, P.-H.C., Liu, Y. & Peng, L. (2019). How to develop machine learning models for healthcare. Nature Materials, 18 (5). pp. 410–414. DOI: 10.1038/s41563-019-0345-0.
6. Corbett, T., Cummings, A., Calman, L., Farrington, N., Fenerty, V., Foster, C., Richardson, A., Wiseman, T. & Bridges, J. (2020). Self-management in older people living with cancer and multi-morbidity: A systematic review and synthesis of qualitative studies. Psycho-Oncology, 29 (10). pp. 1452–1463. DOI: 10.1002/pon.5453.
7. Hang, Choi & Kim (2019). A Novel EMR Integrity Management Based on a Medical Blockchain Platform in Hospital. Electronics, 8 (4). pp. 467. DOI: 10.3390/electronics8040467.
8. Houfani, D., Slatnia, S., Kazar, O., Zerhouni, N., Merizig, A. & Saouli, H. (2020). Machine Learning Techniques for Breast Cancer Diagnosis: Literature Review. In: pp. 247–254. DOI: 10.1007/978-3-030-36664-3_28.
9. Jain, A. (2019). A Survey on Feature Selection for Chronic Diseases Classification. 6 (12). pp. 960–966.
10. Jain, D. & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19 (3). pp. 179–189. DOI: 10.1016/j.eij.2018.03.002.
11. Kaushal, C. & Singla, A. (2021). Analysis of Breast Cancer for Histological Dataset Based on Different Feature Extraction and Classification Algorithms. In: pp. 821–833. DOI: 10.1007/978-981-15-5113-0_69.
12. Kuipers, S.J., Nieboer, A.P. & Cramm, J.M. (2020). Views of patients with multi-morbidity on what is important for patient-centered care in the primary care setting. BMC Family Practice, 21 (1). pp. 71. DOI: 10.1186/s12875-020-01144-7.
13. LaPierre, N., Ju, C.J.-T., Zhou, G. & Wang, W. (2019). MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods, 166. pp. 74–82. DOI: 10.1016/j.ymeth.2019.03.003.
14. Learning, U. machine (2016). *Breast Cancer Wisconsin (Diagnostic) Data Set*. 2016.
15. Li, F.-Q., Wang, S.-L. & Liu, G.-S. (2019). A Bayesian Possibilistic C-Means clustering approach for cervical cancer screening. Information Sciences, 501. pp. 495–510. DOI: 10.1016/j.ins.2019.05.089.

16. McGetrick, J.A., Raine, K.D., Wild, T.C. & Nykiforuk, C.I.J. (2019). Advancing Strategies for Agenda Setting by Health Policy Coalitions: A Network Analysis of the Canadian Chronic Disease Prevention Survey. Health Communication, 34 (11). pp. 1303–1312. DOI: 10.1080/10410236.2018.1484267.
17. Singh, B.K. (2019). Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm. Biocybernetics and Biomedical Engineering, 39 (2). pp. 393–409. DOI: 10.1016/j.bbe.2019.03.001.
18. Singh, S. (2017). diabetes.csv. 2017.
19. Sporinova, B., Manns, B., Tonelli, M., Hemmelgarn, B., MacMaster, F., Mitchell, N., Au, F., Ma, Z., Weaver, R. & Quinn, A. (2019). Association of Mental Health Disorders With Health Care Utilization and Costs Among Adults With Chronic Disease. JAMA Network Open, 2 (8). pp. e199910. DOI: 10.1001/jamanetworkopen.2019.9910.
20. Tseng, Y.-J., Huang, C.-E., Wen, C.-N., Lai, P.-Y., Wu, M.-H., Sun, Y.-C., Wang, H.-Y. & Lu, J.-J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. International Journal of Medical Informatics, DOI: 10.1016/j.ijmedinf.2019.05.003.
21. Zhang, J., Wang, S., Courteau, J., Chen, L., Guo, G. & Vanasse, A. (2019). Feature-weighted survival learning machine for COPD failure prediction. Artificial Intelligence in Medicine, 96. pp. 68–79. DOI: 10.1016/j.artmed.2019.01.003.