
Patrick Bujok  
Department of Business  
Development and Technology  
Aarhus University  
Herning, Denmark  
pb18+IEEE@mailbox.org

Maria Jensen  
Department of Business  
Development and Technology  
Aarhus University  
Herning, Denmark  
mariajensen712@gmail.com

Steffen M. Larsen  
Department of Business  
Development and Technology  
Aarhus University  
Herning, Denmark  
steffen.molgaard@gmail.com

Robert A. Alphinas  
Department of Business  
Development and Technology  
Aarhus University  
Herning, Denmark  
roal@btech.au.dk

Abstract—Machine Learning applications provide a promising method to support clinical practitioners in Breast Cancer (BC) detection. Currently, Fine Needle Aspiration (FNA) is a commonly applied diagnostic method for BC tumors, which, however, is associated with ominous false negative classifications. For this purpose, the present study explores Artificial Neural Networks (ANNs) with the aim of outperforming clinical practices via FNA in classifying benign or malignant BC cases with regard to an improved accuracy and reduced False Negative Rate (FNR) using the Breast Cancer Wisconsin (Diagnostic) Dataset (WDBC). The findings reveal that a dense ANN with a single hidden layer including 15 neurons can reach a testing accuracy of 98.60% and a FNR of 0% on a scaled dataset. In combination with several introduced improvement measures, a high degree of generalizability is associated with the model under consideration of the relatively small dataset. As a result, this model outperforms not only clinical practitioners but also 72 classifiers from the recent literature.

Index Terms—Machine Learning, Breast Cancer, Artificial Neural Network, Classification, WDBC

I. INTRODUCTION

Breast Cancer (BC) is considered as a major threat in medical care [1], which almost solely occurs in women [2]. In 2020, BC accounts for the highest number of new cancer cases, approximately 2.3 million (11.7%), and for the number one of cancer-related mortality rates for women with 685,000 (15.5%) casualties [2]. The invasive BC carcinoma are the origin for the cause of death. A common detection practice is to identify tumors as an indicator of breast carcinoma [3]. These tumors are classified into two types: normal and abnormal of which the abnormal follows the classification of nonharmful (benign) and cancerous (malignant). While benign tumors have a slow growth rate and do not attack other cells nor spread through the body [4], [5], malignant tumors can spread across, for example, muscles, skin, and bloodstream and divide into new, secondary tumors; the so-called metastasis. The identification of benign or malignant requires active diagnostic techniques by clinicians to determine the respective tumor class [5]. Detecting malignant tumors in an early stage is vital for the patient because it reduces the odds of death [5], [6].

In medical practice, the first procedure of BC detection takes place via mammography, which is consulted as a standardized application to screen for BC [7]. Once the mammography analysis identifies an anomaly, a Fine Needle Aspiration (FNA) cytology is commonly performed to provide a diagnosis with a higher accuracy. The FNA procedure typically produces a reduced amount of False Positive (FP) diagnoses compared to a mammography [8]. However, even specialists require strong efforts to provide a correct diagnosis [3], [9].

In this context, Mendoza et al. [8] conducted a review of eight studies about FNA diagnoses with 382 to 1,583 total number of cases of which the False Negative (FN) results range between 1.2% and 10%. In contrast, FPs contribute to 0% to 2% of the total number of cases [8]. In a Japanese study on 10,890 individuals, 3.31% FNs and 0.25% FPs were identified in FNA biopsy diagnoses. The accuracy rate accounts for 88.0% [10]. Two more recent studies present the following findings: An investigation by Aker et al. [11] found a False Negative Rate (FNR) of 1.4% and a FP rate of 2.6% in 4,860 cases [11]. In an investigation of 698 cases, Kazi et al. [9] identified two patients (0.29%) that were falsely classified as having a benign tumor (FN), while no FP were identified [9]. Consequently, misdiagnoses still occur after FNA in particular in FN cases. The findings by Mendoza et al. [8] and Yamaguchi et al. [10] refer to data almost a decade ago, providing a weighted arithmetic mean of 3.7% for all FN cases with regard to the total number of individual cases of both studies. In contrast, the recent studies by Aker et al. [11] and Kazi et al. [9] reveal a weighted arithmetic mean of 1.3% for FN diagnoses.

In comparison to these clinical BC detection practices, Machine Learning (ML) has been proven as a promising tool for classifying benign and malignant tumors providing higher accuracy than physicians. At the same time, these techniques decrease the rate of misdiagnosis, which can account for up to almost 20% of diagnostic cases [12]. Nonetheless, misclassification is still a threat leading to harmful consequences for patients. On the one hand, a FN diagnose can lead to a
dangerous stage since an unidentified malignant tumor can grow until the next doctor’s visit [13]. On the other hand, FP classification is associated with severe negative physical and mental consequences for the misdiagnosed patients [13]. As a result of these misclassifications, ML models require special attention to class imbalance by applying various approaches such as adequate sampling strategies, evolution metrics, or classification algorithms [12]–[15].

Consequently, this work seeks to outperform clinical practices via FNA in classifying benign and malignant BC tumors by applying ML techniques on the Breast Cancer Wisconsin (Diagnostic) Dataset (WDBC) [16] with focus on providing a high model accuracy and minimizing misclassification in form of a reduced FNR. For a quantitative comparison, the FNR of 3.7% in clinical FNA was opted for as maximum baseline due to the high number in total cases and individual samples in the literature. The FNR of 1.3%, in contrast, refers to recent findings, which are less representative because of the significantly lower amount of cases and samples. Besides the maximum FNR, an overall minimum accuracy of 95% is targeted. On the basis of these criteria, the following research questions is investigated: Can an Artificial Neural Network (ANN) be trained to achieve a FNR below 3.7% and an accuracy of at least 95% while providing a high degree of generalizability in classifying benign or malignant BC cases using the WDBC?

II. METHODOLOGY

The methodological strategy of this research follows the approach by Géron [17] in which an end-to-end ML project is executed according to the following steps: (a) Look at the big picture, (b) get the data, (c) discover and visualize the data to gain insights, (d) prepare the data for algorithms, (e) select a model and train it, (f) fine-tune the model, (g) present the solution, as well as (h) launch, monitor, and maintain the system. The last step is excluded from this research.

Step (a) is covered in Section I by framing the problem into the research question. Regarding step (b) and (c), the data are acquired, examined, and visualized beyond the here presented documentation. Subsequently, step (d) is executed by preparing the WDBC for the algorithm and by preprocessing the dataset using the techniques data splitting and data scaling. In step (e), the ANN is created by testing various model architectures manually, starting with a single layer perceptron, followed by one and two hidden layers. With regard to the number of neurons per hidden layer, the strategy foresees a decreasing number of neurons from input to output layer. Based on these constraints, the preprocessed data are used to train the different models considering the Adaptive Moment Estimation (Adam) optimizer. Afterward, the best performing model architecture is chosen and is subjected to fine-tuning, step (f). Here, the model is trained multiple times while the learning rate is manually adjusted for each training session. The aim is to identify a suitable learning rate that converges the model’s error to a minimum resulting in an robust and generalizable model [18]. As a last step (g), the performance of the final model is evaluated on the metrics FNR and accuracy by comparing the outcome with the given threshold. Despite a minimum of randomness in the partitioning of data and during training phase, the outcome of this particular research can be reproduced with the applied program code. However, executing the code may not reproduce the exact same numerical results, yet reproducibility is given to answer the research question likewise.

III. THEORY

A. Model Evaluation

The evaluation of the model is achieved by means of the metrics accuracy, precision, recall and FNR. The confusion matrix in Table I provides an overview of the model performance consisting of True Positives (TPs), True Negatives (TNs), FPs, and FNs to which the previous mentioned metrics can be calculated with (1) to (4).

\[
\begin{align*}
\text{Accuracy} & = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1) \\
\text{Precision} & = \frac{TP}{(TP + FP)} \quad (2) \\
\text{Recall} & = \frac{TP}{(TP + FN)} \quad (3) \\
\text{FNR} & = 1 - \text{Recall} = \frac{FN}{(FN + TP)} \quad (4)
\end{align*}
\]

B. Dense Artificial Neural Network

Inspired by the brain’s architecture and its usage of networks of biological neurons [19], [20], a dense neural network consist of three different types of layers: (i) the input layer, which consist of several pass-through neurons passing on fed information to the next layer [19], (ii) the hidden layer, which consist of a number of artificial neurons connected with a weight to the following layer [20], and (iii) the output layer in which the number of neurons is determined by the problem at hand. Data are computed and flows through the neurons in each layer. When a threshold value is reached in a neuron, the information is feedforwarded through the connections from the input layer to the output layer (Fig. 1). Here, the strength of the connections between each neuron to the following layer is determined by the weights.

Learning is achieved by means of backpropagation. This type of algorithm improves the model performance by adjusting the weighted connection between each neuron in the subsequent layer, according to:

\[
f(x) = \sum (w \cdot x) + b \quad (5)
\]

<table>
<thead>
<tr>
<th>Actual</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>True</td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>FN</td>
</tr>
</tbody>
</table>
A. Dataset and Preprocessing

The WDBC origins from the University of Wisconsin Hospitals and has historically been used to conduct breast tumor diagnoses [16]. The dataset derives from digitized images of a breast biopsy, obtained by means of FNA. In this process, a small gauge needle has been used to extract fluid from a breast lump or mass, which previously has been detected by the patient or in a mammography [25].

All 30 input features are used to estimate whether the tumor is malignant or benign. The WDBC dataset is split into training and testing data with a size of 75% and 25% respectively. The training dataset consists of 177 malignant cases out of 249 observations while the testing dataset consists of 35 malignant cases out of 108 observations. The dataset is scaled by the help of StandardScaler, where $x$ is the observation, $u$ is the sample mean, and $s$ is the sample variance, as follows:

$$z = (x - u)/s$$

B. Dense Artificial Neural Network Design

The ANN for classifying benign and malignant cases on WDBC is built with the help of the TensorFlow library [26]. The ANN architecture is instantiated as a sequential model, which is a simple stack of layers connected sequentially. Fig. 2 shows the ANN architecture design consisting of an input layer with 30 neurons, a dense hidden layer with 15 neurons, and a dense output layer with a single neuron for binary classification.

The ReLU activation function is applied for the dense hidden layer and the sigmoid activation function is used for the output layer. During the training phase, the model uses 10% of the training data for validation and initializes EarlyStopping as a callback parameter.
V. RESULTS AND DISCUSSION

A. Accuracy and False Negative Rate

As visualized in Fig. 3, EarlyStopping is instantiated after 38 epochs, rendering a fully trained model with its confusion matrices in Table II and Table III. Based on these matrices, it can be derived that the created model has an accuracy of 98.83% and a FNR of 2.82% in the training phase. During the testing phase, the model achieves an accuracy of 98.60% and a FNR of 0%, as summarized in Table IV. From a quantitative perspective, this work has proven that an ANN can outperform clinical practices via FNA in classifying benign or malignant tumors by achieving an accuracy of over 95% and a FNR of below 3.7%.

B. Generalizability

An algorithm with a high accuracy commonly arises the concern of overfitting the training dataset. However, the accuracy of the presented ANN on the training (98.83%), validation (97.67%), and testing (98.60%) data are almost identical, which indicates that the high accuracy is not caused by the noise in the training data. Moreover, according to Mangasarian and Wolberg [27], the high accuracy can be obtained due to the fundamental distinction in the characteristic of the two tumor classes. The researchers show that the presence of a significantly higher magnitude of multiple feature values indicates an increased probability of malignancy. This factor additionally diminishes the risk of an overfitting model [27].

Another aspect, which influences the generalizability of the model on unseen data, is the number of available observations. Due to the division of the dataset into 75% training data and 25% testing data, a comparatively small share of 35 malignant observations is present in the testing phase. Hence, the size of the dataset is insufficient to perform a proper data split into training, validation, and testing data. To guarantee a representative partition of validation data, the ANN uses 10% of the training data for validation during the training phase. This common technique on small datasets increases the likelihood of a generalizable model. After all, it may be inferred that the introduced measures to the model result in a high degree of generalizability under the consideration of the limited availability of data.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>CONFUSION MATRIX OF TRAINING PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Actual</td>
</tr>
<tr>
<td>Benign</td>
<td>TP = 172 (97.18%)</td>
</tr>
<tr>
<td>Malignant</td>
<td>FN = 0 (2.82%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>CONFUSION MATRIX OF TESTING PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Actual</td>
</tr>
<tr>
<td>Benign</td>
<td>TP = 35 (100%)</td>
</tr>
<tr>
<td>Malignant</td>
<td>FN = 0 (0%)</td>
</tr>
</tbody>
</table>

C. Literature Comparison

In the scientific literature, several techniques have been frequently applied on the WDBC. In order to guarantee a higher degree of comparability, the following discussion solely focuses on literature findings that used the identical dataset. In this sense, the literature was reviewed and the testing performance of 72 classifiers was collected [28]–[38]. The accuracy ranges between 81.37% and 99.04%. The FNR was determined using (4) and its values range between 0% and 49.09%. Considering the given threshold, 19 out of 72 (26%) classifiers outperform clinical practices. Table V lists these 19 algorithms and compares its performance with the ANN of this study in terms of the training and testing phase. Hereby, the ANN is:

1) ▲ superior in accuracy and FNR,
2) ▼ inferior in accuracy and FNR, or
3) ▼ outperforming either accuracy or FNR.

It is evident that the ANN is superior in almost all cases with regard to the testing phase except the Multilayer Perceptron (MLP) by Agarap [28]. Comparing the ANN training phase with the given testing phase from the literature, then every classifier is inferior to the ANN. Furthermore, 5 out of 19 (26%) algorithms are even outperformed by the ANN in terms of both accuracy and FNR. However, a solely quantitative comparison of these indicators has to be taken with caution due to the fact that various factors influence the results including distinct classification techniques, the separation in training and testing data, and a certain degree of randomness in the execution. For that reason, the following qualitative discussion focuses on the high performing classifiers in the literature with FNRs of zero or close to zero.

Agarap [28] investigated the performance of six algorithms of which four fulfil the given criteria regarding accuracy

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>MODEL PERFORMANCE OF TRAINING AND TESTING PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Training</td>
<td>98.83%</td>
</tr>
<tr>
<td>Testing</td>
<td>98.60%</td>
</tr>
</tbody>
</table>
Using Independent Component Analysis (ICA) in combination with four algorithms. A one-dimensional feature vector reduced by ICA was compared to the full dataset of 30 features. A 10-fold cross-validation (CV) with a 80% and 20% split in training and testing phase respectively was applied on both feature sets. The ANN by Mert et al. [33] outperforms clinical practices using the full dataset, while the ICA method does not exceed the given threshold. Nonetheless, the ANN from this study demonstrates its superiority over its counterpart with regard to the testing phase. Furthermore, it should be noted that the FNR was not further discussed by Mert et al. [33].

After all, a high number of findings perform better than clinical practitioners. Despite this high performance of ML algorithms, the critical case of falsely negative diagnosed patients is not closer discussed in various studies, even though FNR is a vital parameter in the context of BC detection. A strong model with high accuracy should be still considered as critical when the FNR does not provide a desired magnitude. A high accuracy with a high FNR should be avoided compared to a reduced accuracy with a very low FNR.

VI. CONCLUSION

In this study, the WDBC is used to determine if ML algorithms can outperform clinical practitioners in classifying benign and malignant BC cases on observations extracted via FNA. Following a literature review on BC detection via FNA in health care, a threshold of a FNR below 3.7% and an accuracy above 95% is determined. The aim is to outperform the defined threshold with a trained model. For this purpose, a simple dense ANN has been developed and trained using the WDBC. As a result, the model outperforms clinical practitioners on both training and testing data. In detail, the ANN achieved an accuracy of 98.83% and a FNR of 2.82% in the training phase and an accuracy of 98.60% and a FNR of 0% in the testing phase. The high accuracy in the training, validation, and testing phase is an indicator for a model that is neither underfitting nor overfitting to the training data. Consequently, the degree of generalizability can be considered as high with regard to the limited dataset.

This ANN model has been compared with in total 72 classifiers from the literature on the basis of the WDBC. The performance of these models is given by the accuracy and FNR, which are solely based on the testing results. The comparison of the testing phase reveals that firstly only 19 (26%) classifiers outperform clinical practices based on the given thresholds and secondly all of these models are outperformed by the presented ANN model from this study. In addition, the generated ANN is superior to 5 out of the top 19 (26%) models during the testing and training phase in terms of both accuracy and FNR. Furthermore, a close analysis of the top performing algorithms shows that the essential reduction of the FNR is neither seen as critical nor is it considered as part of the research purpose. In conclusion, this study provides a superior dense ANN in BC detection compared to clinical practices and existing literature findings. Moreover, this work fills the gap of the missing focus on FNR reduction using

<table>
<thead>
<tr>
<th>Reference / Classifier</th>
<th>Literature ANN*</th>
<th>Train (Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression (30% testing, features = 30)</td>
<td>96.09</td>
<td>0</td>
</tr>
<tr>
<td>MLP (30% testing, features = 30)</td>
<td>99.04</td>
<td>0.79</td>
</tr>
<tr>
<td>Softmax Regression (50% testing, features = 30)</td>
<td>97.66</td>
<td>0</td>
</tr>
<tr>
<td>SVM (RBF, 30% testing, features = 30)</td>
<td>96.09</td>
<td>2.47</td>
</tr>
<tr>
<td>Bagiri et al. [29]</td>
<td>k-RNN (60% testing, features = 3)</td>
<td>98.1</td>
</tr>
<tr>
<td>Esmans and Parkin [30]</td>
<td>AP-AMIBFA (10-CV, epoch = 100)</td>
<td>98.46</td>
</tr>
<tr>
<td>Mert et al. [32]</td>
<td>AMIBFA (10-CV, epoch = 100)</td>
<td>98.21</td>
</tr>
<tr>
<td>Mert et al. [32]</td>
<td>AP-BFA (10-CV, epoch = 100)</td>
<td>98.54</td>
</tr>
<tr>
<td>Mert et al. [33]</td>
<td>BFA (10-CV, epoch = 100)</td>
<td>98.17</td>
</tr>
<tr>
<td>Kadam et al. [31]</td>
<td>SSAE-SM (10-CV, epoch = 50 to 600)</td>
<td>98.243</td>
</tr>
<tr>
<td>Mert et al. [32]</td>
<td>FE-SSA-SM (10-CV, epoch = 50 to 600)</td>
<td>98.594</td>
</tr>
<tr>
<td>Mert et al. [33]</td>
<td>PSN (10-CV, features = 30)</td>
<td>96.31</td>
</tr>
<tr>
<td>Mert et al. [33]</td>
<td>PSN (leave-one-out CV, features = 30)</td>
<td>97.61</td>
</tr>
<tr>
<td>Mert et al. [33]</td>
<td>ANN (10-CV, features = 30)</td>
<td>97.54</td>
</tr>
<tr>
<td>Obaid et al. [34]</td>
<td>SVM (13-CV, linear Kernel)</td>
<td>97.9</td>
</tr>
<tr>
<td>Pohle et al. [35]</td>
<td>SVM (10-CV, linear Kernel)</td>
<td>98.07</td>
</tr>
<tr>
<td>Pohle et al. [35]</td>
<td>k-NN (10-CV, k = 3)</td>
<td>97.36</td>
</tr>
<tr>
<td>Pohle et al. [35]</td>
<td>J48 (10-CV)</td>
<td>98.07</td>
</tr>
<tr>
<td>Pohle et al. [35]</td>
<td>Random Forest (10-CV)</td>
<td>97.77</td>
</tr>
<tr>
<td>Pohle et al. [35]</td>
<td>MLP (10-CV)</td>
<td>98.41</td>
</tr>
</tbody>
</table>

* Classifier from this study

a ANN is superior (▲), inferior (▼), or outperforming either accuracy or FNR (▼▼)

and FNR. The data for the analysis was standardized and divided in 70% training data and 30% testing data. It is shown that the MLP reveals best performance with a very high accuracy of over 99% and a FNR of 0.79%. Quantitatively, the MLP outperforms the ANN training results in terms of both accuracy and FNR, while the ANN is superior in FNR with respect to the testing result.

By initiating a qualitative comparison between the MLP and the ANN from this study, it is shown that both neural networks use ReLU as activation function and binary cross-entropy as loss function. Despite these similarities, the architecture of the networks differs fundamentally. The applied structure in the MLP consists of three hidden layers with 500 nodes per layer resulting in a 500-500-500 architecture with 512,896 trainable parameters, which were trained with 3,000 epochs. In contrast, the ANN holds one hidden layer with 15 neurons and 481 trainable parameters, which were trained with 38 epochs. Due to the significantly larger architectural size of the MLP, it can be assumed that the ANN executes predictions with a higher efficiency considering the reduced number of trainable parameters by the factor >1,000. Moreover, it can be presumed that the MLP operates a network, which potentially stores the information of the 569 data points rather than providing generalizable predictions. For these reasons, the ANN is considered as a superior performing classifier. Lastly, it has to be mentioned that the FNR was given by Agarap [28] but not closer discussed.

Mert et al. [33] studied the properties of feature reduction using Independent Component Analysis (ICA) in combination with four algorithms. A one-dimensional feature vector reduced by ICA was compared to the full dataset of 30 features. A 10-fold cross-validation (CV) with a 80% and 20% split in training and testing phase respectively was applied on both feature sets. The ANN by Mert et al. [33] outperforms clinical practices using the full dataset, while the ICA method does not exceed the given threshold. Nonetheless, the ANN from this study demonstrates its superiority over its counterpart with regard to the testing phase. Furthermore, it should be noted that the FNR was not further discussed by Mert et al. [33].

After all, a high number of findings perform better than clinical practitioners. Despite this high performance of ML algorithms, the critical case of falsely negative diagnosed patients is not closer discussed in various studies, even though FNR is a vital parameter in the context of BC detection. A strong model with high accuracy should be still considered as critical when the FNR does not provide a desired magnitude. A high accuracy with a high FNR should be avoided compared to a reduced accuracy with a very low FNR.
ML application in health care. Further research and larger data samples are required to validate the findings before the implementation phase into practice can be introduced.

REFERENCES


