UDC 004.67: 612.113 **doi: 10.32620/reks.2024.4.08**

Gennady CHUIKO, Denys HONCHAROV

Petro Mohyla Black Sea National University, Mykolaiv, Ukraine

DIMENSIONALITY CUTBACK AND DEEP LEARNING ALGORITHMS EFFICACY AS TO THE BREAST CANCER DIAGNOSTIC DATASET

Breast cancer is a significant threat because it is the most frequently diagnosed form of cancer and one of the leading causes of mortality among women. Early diagnosis and timely treatment are crucial for saving lives and reducing treatment costs. Various medical imaging techniques, such as mammography, computed tomography, histopathology, and ultrasound, are contemporary approaches for detecting and classifying breast cancer. Machine learning professionals prefer Deep Learning algorithms when analyzing substantial medical imaging data. However, the application of deep learning-based diagnostic methods in clinical practice is limited despite their potential effectiveness. Deep Learning methods are complex and opaque; however, their effectiveness can help balance these challenges. ***The research subjects.*** *Deep Learning algorithms implemented in WEKA software and their efficacy on the Wisconsin Breast Cancer dataset.* ***Objective****. Significant cutback of the dataset's dimensionality without losing the predictive power.* ***Methods****. Computer experiments in the WEKA medium provide preprocessing, supervised, and unsupervised Deep Learning for full and reduced datasets with estimations of their efficacy.* ***Results****. Triple sequential filtering notably reduced the dimensionality of the initial dataset: from 30 attributes up to four. Unexpectedly, all three Deep Learning classifiers implemented in WEKA (Dl4jMlp, Multilayer Perceptron, and Voted Perceptron) showed the statistically same performance. In addition, the performance was statistically the same for full and reduced datasets. For example, the percentage of correctly classified instances was in range (95.9-97.7) with a standard deviation of less than 2.5 %. Two clustering algorithms that use neurons (Self Organized Map, SOM, and Learning Vector Quantization, LVQ) have also shown similar results. The two clusters in all datasets are not well separated, but they accurately represent both preassigned classes, with the Fowlkes–Mallow indexes (FMI) ranging from 0.81 to 0.99.* ***Conclusion****. The results indicate that the dimensionality of the Wisconsin Breast Cancer dataset, which is increasingly becoming the "gold standard" for diagnosing Malignant-Benign tumors, can be significantly reduced without losing predictive power. The Deep Learning algorithms in WEKA deliver excellent performance for both supervised and unsupervised learning, regardless of whether dealing with full or reduced datasets.*

Keywords: *breast cancer; Deep Learning algorithms; WEKA; Wisconsin Breast Cancer dataset; diagnosing Malignant-Benign tumors.*

1. Introduction

1.1. Motivation

The incidence of breast cancer (BC) is increasing in Ukraine, with mortality rates similar to those in Europe. One of the profound reasons for this is the diagnosis of a disease that is too late. Currently, medical observations detect less than a third of BC cases, and this fraction is decreasing steadily [1]. The up-to-date homeland screening mammography studies included in the medical guarantee program have an unacceptably low coverage of women from the target groups (only 3.7%). Aside from this, there is an alarming lack of medical awareness and activity among our women, resulting in only about 17% of them getting mammograms even when referred for the procedure [1].

Research conducted in Europe showed that implementation of mammographic screening decreases breast cancer mortality among women. Under optimal coverage conditions, it is estimated that 23% more breast cancer deaths could be prevented in Eastern Europe, compared to 21% in Western Europe, 15% in Southern Europe, and 9% in Northern Europe [2].

Early and effective detection of this disease significantly increases the survival rate and reduces treatment costs [3]. In recent decades, machine learning (ML) and Deep Learning (DL) have emerged as valuable tools in data-driven decision-making, for example, within resource management [4]. Besides, they are recognized as contemporary methods for the early diagnosis of breast cancer (BC) [3].

1.2. State of the art

Several well-known datasets related to female breast cancer have been used in machine learning (ML). A notable dataset was obtained from the Institute of

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Oncology at the University Medical Center in Ljubljana, which was first made available in 1988 [5]. The proposed dataset contains 10 attributes and contains 286 instances. It includes two binary classes: "no-recurrence events," with 201 cases, and "recurrence events," with 85 instances. This dataset is considered noisy and demonstrated relatively low performance. However, it was recently partially cleaned to reduce noise and improve performance [6].

The Wisconsin Breast Cancer Dataset (WBCD), which has been in use since 1995, comprises 30 attributes and 569 instances [7]. This class includes two imbalanced tumor classes: 212 malignant and 357 benign. The dataset focuses on the geometric parameters of tumors identified through mammography screening images. The newer BreakHis database ([8] and [9]) can expand on this dataset by providing additional information on biopsy, tumor class, tumor type, patient ID, and magnification factor. It is worth noting that the extension of WBCD requires some caution because the initial dataset [7] was already sufficiently bulky.

The use of Deep Learning techniques is effective for detecting breast cancer, enabling early diagnosis, and increasing patient survival. First, Deep Learning (DL) requires less human intervention for feature extraction than classical Machine Learning (ML) techniques [10]. Second, the DL methods are suitable for bulky datasets, like WBCD, although they require more machine resources. Finally, Deep Learning has become a standard tool for breast cancer detection. For instance, DL methods can diagnose breast cancer up to 12 months earlier than conventional clinical procedures [11].

A long time ago, we observed that the predictive power of any classifier initially increased with the number of dimensions (number of attributes). However, after reaching some dimension size, the performance degrades using a fixed-size training set. This effect is known as the "curse of dimensionality" or the Hughes phenomenon [12].

Machine learning, particularly DL, cannot avoid this problem. Volumetric WBCD, with its 30 attributes, certainly needs the correct lowering of dimensionality. Such attempts are being made using WEKA – a Java-based environment for ML [13]. The spread insight that DL can achieve high performance regardless of the dimensionality of the feature space is, to be sure, nothing more than a harmful illusion.

1.3. Objectives and the Approach

This research aimed to improve WBCD's predictivity power and enhance clinical usability using a few DL algorithms implemented in WEKA. Sundry tasks will be performed to achieve this goal:

1) The dataset should be thoroughly preprocessed, including standardizing numeric attributes, optimizing their selection, and using principal component analysis to reduce dimensionality.

2) WEKA comparable experiment with three DL classifiers and three datasets (complete and two gradually reduced ones), the design of that includes tuning of hyperparameters for DL classifiers (supervised deep learning).

3) Unsupervised Deep Learning and collating the efficacy of two DL clustering algorithms within the Knowledge Flow module of WEKA in work with the three above datasets.

The list of tasks determines the research approach. Thus, preprocessing and reduction of the initial WBCD are considered in section "2. Materials and methods." Supervised Deep Learning experiments and clustering will be presented in the following two subsections of section "3. Results." Sections for Discussion and Conclusions will be on the traditional places.

2. Materials and methods of research

2.1. Data are "Materials" within Machine Learning

Therefore, we begin by describing the WBCD dataset. Ten valid characteristics were calculated for each cell nucleus extracted from the mammographic images [7]:

1) radius (mean of distances from the center to points on the perimeter)
2) texture (standard deviation of gray-scale values)
3) perimeter
4) area
5) smoothness (local variation in radius lengths)
6) compactness (perimeter^2 / area)
7) concavity (severity of concave portions of the contour)
8) concave points (number of concave portions of the contour)
9) symmetry
10) fractal dimension ("coastline approximation")

The WBCD dataset foresees three attributes: mean, standard deviation, and "worst" (extreme) value for each attribute. "Worst values" are factually outliers from a statistical perspective [13]. As a result, the number of numeric features was increased to 30. All numeric features (attributes) are continuous and have no missing values.

The single categorial feature is the nominal class: benign (357) or malignant (212) diagnosis, without missing values. Perhaps it is a "trade-off matter" to consider this class nearly balanced or inversely. In this article, we

selected imbalance as our insight. WBCD provided sufficient precision for neural network classifiers in the range (0.865 - 0.9597) [7].

2.2. Methods

First, all numeric attributes were standardized with an attribute filter incorporated into WEKA

weka.filters.unsupervised.attribute.Standardize

As a result, all numeric attributes have zero mean values and standard deviations of unity. Standardization assumes that attributes have a Gaussian (bell curve) distribution. This does not have to be strictly proven; however, this technique is beneficial if the attribute distribution is closer to Gaussian. WEKA automatically builds histograms for all attributes. A simple visual analysis of these histograms demonstrates that most of the features (attributes) of WBCD have distributions that are close enough to Gaussian ones. Thus, our first dataset (ds1) for the following experiment included 30 filtered (standardized) numerical attributes and one nominal class.

The second reduced dataset (ds2) was obtained by further filtering ds1 through an attribute selection filter (CfsSubsetEval). The configuration of this filter was set using the following Java-line:

weka.filters.supervised.attribute.AttributeSelection -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5".

This dataset (ds2) contains only 11 standardized attributes against 30 in ds1 and one nominal class. Among these 11 attributes of the reduced dataset are six of the "worst" type, three of "mean," and two of "se" following [13].

Eleven attributes of ds2 are far superior to the 30 in ds1, yet still considered "excessive." Therefore, we should implement an additional filter to condense the 11 attributes into a lower-dimensional space. This filter can be a principal component filter with the following configuration:

weka.filters.unsupervised.attribute.PrincipalComponents -R 0.91 -A 11 -M -1.

This filter reduces ds2 from 11 to four principal components, capturing 91% of the total variance. It is always tempting to reduce the number of principal components even further, perhaps to two. However, doing so would mean accepting a smaller share of the total variance coverage; therefore, there is a "trade-off" to consider. The resulting dataset, which was triple-filtered and reduced to four attributes (ds3), can be regarded as "oversimplified." However, we will wait until the comparative analysis results are ready.

WEKA includes several DL algorithms: Multi-Layer Perceptron (MLP) [14], Voted Perceptron [15], and the newer Dl4jMlp [16]. Although all of them are based on Rosenblatt's prototype [14], they still have different tuning options. In our experiment, these tunings (configurations) are presented in (Table 1).

Table 1

Classifier's configurations

Algorithm	**Configuration**
MLP	*weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a*
Voted Perceptron	*weka.classifiers.functions.VotedPerceptron -I 1 -E 1.0 -S 1 -M 10000*
Dl4jMlp	*weka.classifiers.functions.Dl4jMlpClassifier -S 1 -cache-mode MEMORY -early-stopping "weka.dl4j.earlystopping.EarlyStopping -maxEpochsNoImprovement 0 -valPercentage 0.0" -normalization "Standardize training data" -iterator "weka.dl4j.iterators.instance.DefaultInstanceIterator -bs 1" -iteration-listener "weka.dl4j.listener.EpochListener -eval true -n 5" -layer "weka.dl4j.layers.OutputLayer -lossFn \"weka.dl4j.lossfunctions.LossMCXENT \" -nOut 2 -activation \"weka.dl4j.activations.ActivationSoftmax \" -name \"Output layer\"" -logConfig "weka.core.LogConfiguration -append true -dl4jLogLevel WARN -logFile C:\\Users\\master\\wekafiles\\wekaDeeplearning4j.log -nd4jLogLevel INFO -wekaDl4jLogLevel INFO" -config "weka.dl4j.NeuralNetConfiguration -biasInit 0.0 -biasUpdater \"weka.dl4j.updater.Sgd -lr 0.001 -lrSchedule \\\"weka.dl4j.schedules.ConstantSchedule -scheduleType EPOCH\\\"\" -dist \"weka.dl4j.distribution.Disabled \" -dropout \"weka.dl4j.dropout.Disabled \" -gradientNormalization None -gradNormThreshold 1.0 -l1 NaN -l2 NaN -minimize -algorithm STOCHASTIC_GRADIENT_DESCENT -updater \"weka.dl4j.updater.Adam -beta1MeanDecay 0.9 -beta2VarDecay 0.999 -epsilon 1.0E-8 -lr 0.001 -lrSchedule \\\"weka.dl4j.schedules.ConstantSchedule -scheduleType EPOCH\\\"\" -weightInit XAVIER -weightNoise \"weka.dl4j.weightnoise.Disabled \"" -numEpochs 10 -numGPUs 1 -averagingFrequency 10 -prefetchSize 24 -queueSize 0 -zooModel "weka.dl4j.zoo.CustomNet -channelsLast false -pretrained NONE"\ "weka.dl4j.weightnoise.Disabled \"" -numEpochs 10 -numGPUs 1 -averagingFrequency 10 -prefetchSize 24 -queueSize 0 -zooModel "weka.dl4j.zoo.CustomNet -channelsLast false -pretrained NONE"*

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

The proposed design provides 10-fold cross-validation for each classification and requires ten repetitions. With three datasets, three classifiers, 10-fold cross-validation, and ten repetitions, we can summarize 900 experimental results. This allows us to conduct specific statistics, hypotheses, and conclusions. The confidence level was set at 0.95 ($p \leq 0.05$). The corrected paired Student's t-test was exploited for statistics hypotheses. The confidence level was set at 0.95 ($p \leq 0.05$). The WEKA Experimenter logs state that such a design demands about 20-21 minutes to execute on a middle-class personal computer.

WEKA offers two DL algorithms for clustering: a) LVQ (Learning Vector Quantization) – an artificial neural network that applies a "winner-take-all" learning-based approach [17]; b) Self Organized Map (SOM, Kohonen's net) [18], which is similar to learning.

In principle, the Experimenter allows us to create an advanced experiment that can collate both clustering algorithms, as described in [19]. Unfortunately, this approach is still not feasible for most clustering algorithms, particularly DL algorithms.

For this reason, we performed clustering of our datasets "manually," ensuring the evaluations of the alignment between preassigned classes and clusters. This means that the clustering mode included "classes-to-clusters" estimations. This clustering mode allows the calculation of complexity matrices for each algorithm and dataset. These matrices are analogous to confusion matrices at the classification level and allow for calculating Fowlkes-Mallows Indexes [20], which are numeric estimates of class-to-cluster congruency.

The configurations of both clustering algorithms were as follows:

weka.clusterers.SelfOrganizingMap -L 1.0 -O 2000 -C 1000 -H 2 -W 1

weka.clusterers.LVQ -L 1.0 -T 1000 -C 2

3. Results and Discussion

3.1. Supervised Deep Learning experiment results

The WEKA experiment described in the previous section allows us to compile performance metrics and their standard deviations for three datasets (ds1, ds2, and ds3) and three Deep Learning algorithms. Table 2 provides an example of this compilation. The percentage of correctly classified instances for each dataset and algorithm (accuracies).

It appears that the table exhibits a slight performance drop across its rows (i.e., ds1->ds2->ds3) and columns. However, this illusion is only an illusion, as none of these "differences" are statistically significant. Thus, all datasets were equally good, and the algorithms were similarly powerful regarding the percentage of instances correctly classified. Furthermore, other well-known performance indices (precision, recall, F-measure, Matthew's correlation coefficient (MCC), and Kappa statistic) are statistically identical across datasets and algorithms.

Table 2

Percentage of correctly classified instances (brackets show the standard deviations)

Datasets	DL Algorithms		
	Dl4jMlp	**MLP**	**Voted Perceptron**
ds1	97.68 (1.80)	96.72 (2.20)	96.45 (2.11)
ds2	97.58 (1.79)	96.64 (1.90)	96.78 (2.14)
ds3	96.40 (2.26)	95.87 (2.47)	95.96 (2.33)

Some exceptions exist regarding the areas under the Receiver Operating Characteristic (ROC) curve and the Precision-Recall Curve (PRC). It is widely recognized that PRC area index values greater than 0.85 indicate a reliable classifier, whereas values approaching 1.0 suggest a perfect classifier. However, it is essential to note that the PRC area values listed in Table 3 for the Voted Perceptron algorithm, although excellent, are statistically significantly worse than those of the other two algorithms. In addition, we found no significant differences between the datasets.

Table 3

The area under the PRC curves

Datasets	Algorithms		
	Dl4jMlp	**MLP**	**Voted Perceptron**
ds1	1.00 (0.01)	0.99 (0.01)	0.97 (0.02) *
ds2	0.99 (0.01)	0.99 (0.01)	0.97 (0.02) *
ds3	0.99 (0.01)	0.99 (0.01)	0.96 (0.02) *

The performances of all three Deep Learning algorithms were evaluated using full and reduced datasets, and all algorithms proved equally effective. For simplicity, we present the classification results for only one algorithm, the Multilayer Perceptron (MLP), and one dataset, which is referred to as ds3. Table 4 presents the confusion matrix and performance metrics for the MLP classifier applied to the most reduced version of the ds3 dataset, where 'M' indicates malignant tumors and 'B' indicates benign tumors (Fig. 1).

Table 4 shows that the Deep Learning classifiers are sufficiently robust for both classes to be roughly alike although the Benign class appears slightly better.

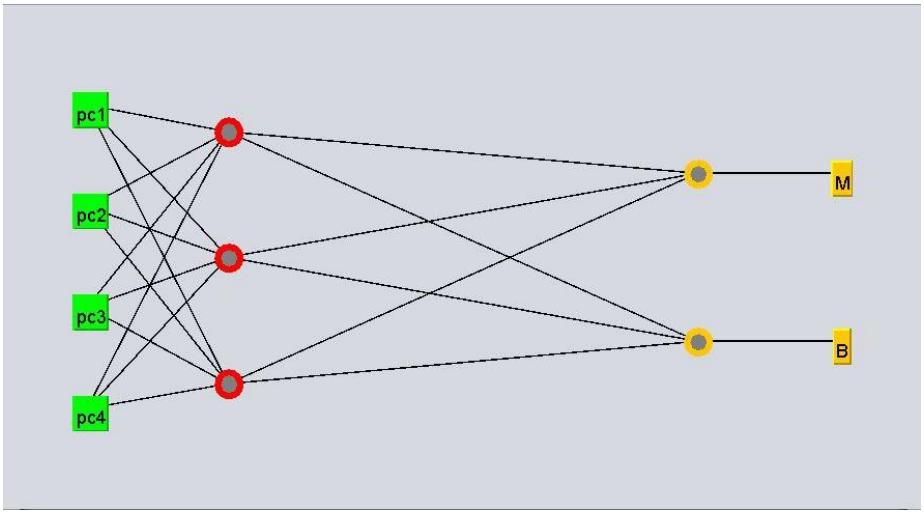

Fig. 1. WEKA screenshot of Multilayer Perceptron (MLP) for reduced dataset ds3; Perceptron contains four neurons (pc1, pc2, pc3, and pc4 meet to the number of principal components) in the input layer, three neurons in the hidden layer, and two output neurons. The MLP architecture is more complex for ds2 and is even more complex for ds1 than for ds3

Table 4

Confusion matrix and performance of MLP algorithm for ds3 dataset

Confusion matrix	Precision	Recall	F-measure	MCC	PRC area	Class
$\begin{pmatrix} 201 & 11 \\ 7 & 350 \end{pmatrix}$	0.966	0.948	0.957	0.932	0.986	M
	0.970	0.980	0.975	0.932	0.994	B

3.2. Unsupervised Deep Learning results

Neural network clustering methods, such as SOM (Self Organized Map), are part of model-based clustering methods. As a typical example, SOM maps a higher-dimensional input space to a lower-dimensional output space, assuming that a specific topology exists in the input data [20]. These methods can effectively separate even overlapping clusters without requiring prior knowledge about the data's topology. However, they have the following drawbacks: a relatively long processing time, and the clustering result is sensitive to the parameters of the selected models.

Figure 2 shows pie charts showing the relative capacities of the clusters obtained by the SOM and LVQ (Learning Vector Quantization) algorithms compared to the relative capacities of the classes. The charts appear similar, but clusters matching malignant tumors have slightly fewer sizes than this class within the dataset and are even more imbalanced.

A higher VRC value indicates that the clusters are dense and well separated although there is no "acceptable" cut-off value. WEKA builds and describes centroids for each cluster; thus, the VRC evaluations are not difficult. The estimated VRC values were 1.78 for the LVQ algorithm and 1.87 for the SOM algorithm. They are undoubtedly low, indicating that clusters are not well separated or overlapped.

Several well-known internal indexes of clustering validity exist [21], including the Variance Ratio Criterion

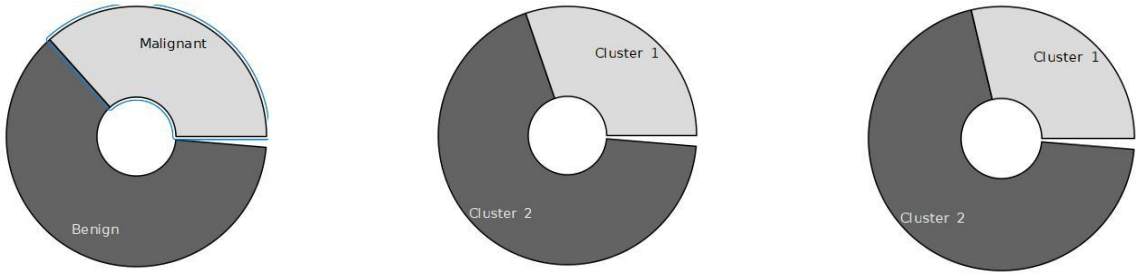

Fig. 2. The pie charts illustrate the class distribution in the entire dataset (ds1, left-hand side) and clusters in the most reduced dataset (ds3, middle, and right-hand side). The middle chart is created using the SOM algorithm, while the right chart uses the LVQ algorithm

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

(VRC), which measures how similar an object is to its cluster (cohesion) compared to other clusters (separation) [22]. Here, cohesion is estimated based on the squared distances from the data points within a cluster to its center, and separation is based on the squared distance between the cluster centroids.

WEKA provides a special clustering mode called "classes to clusters evaluation." This mode yields a matrix that is structurally identical to the confusion matrix. This matrix helps evaluate the congruency between clusters and preassigned classes in the dataset. The numeric estimators for are known as the Fowlkes–Mallow indexes (FMI) [23]. They can be written down as follows:

$$\left.\begin{aligned} \mathrm{FMI}_{\mathrm{positive}} &= \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}} \\ \mathrm{FMI}_{\mathrm{negativw}} &= \frac{\mathrm{TN}}{\sqrt{(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{aligned}\right\} \quad (1)$$

where $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$ are well-known matrix elements for "positive" and "negative" classes.

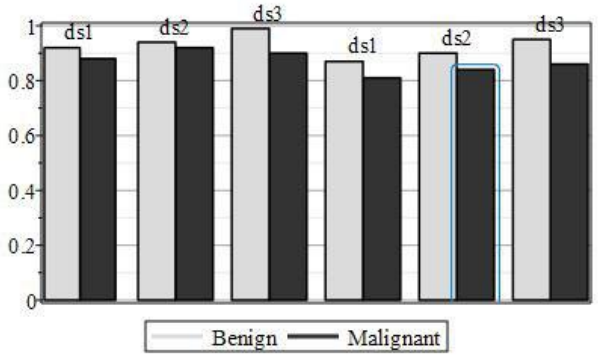

Fig. 3. Fowlkes–Mallow's indexes (FMI) for classes-to-clusters congruency evaluations; three pairs on the left-hand side meet the SOM algorithm, whereas the three pairs of FMI from the right-hand side–the LVQ one

Figure 3 shows the calculated FMI for both clustering algorithms and all datasets. All FMI values were in the range (0.81-0.99), indicating relatively high congruence between clusters and classes. The differences between the benign and malignant tumor classes were not significant. Nevertheless, the systematic differences between them, with the benefit of the first, might be a consequence of the imbalanced data sets.

4. Discussion

Dl4jMlp is a Deep Learning WEKA package that integrates new Deep Learning techniques into the WEKA workbench [24]. This algorithm differs in vast opportunities relative to tuning hyperparameters compared to MLP or Voted Perceptron (see Table 1, for instance). We chose a neural network configuration with a stochastic gradient descent optimization algorithm as follows [24].

The MLP and Voted Perceptron are feedforward artificial neuron networks with at least three layers (see Figure 1). They can effectively separate nonlinearly distinguishable data. The Voted Perceptron differs because it provides more stability to the data size and has weight vectors that offer larger "margins." Surprisingly, the performance of these classifiers was unexpectedly similar when tested on the initial and reduced versions of WBCD. Nevertheless, the performance of all algorithms was excellent regardless of whether the datasets were complete or reduced.

The Unsupervised Deep Learning (clustering) of WBCD and its reduced versions confirms data division into classes. In other words, despite the poor separation of clusters and noise. For example, the Interquartile Range filter (*weka.filters.unsupervised.attribute.The InterquartileRange -R first-last -O 3.0 -E 6.0*) shows the presence of 55 cases that should be recognized as outliers (17 for the Benign class and 38 for the Malignant class, respectively). Thus, emissions comprise approximately 10% of WBCD, which is a challenge to consider as a minor factor. The insensitivity of Deep Learning algorithms to noise is one of the features proposed in this study.

5. Conclusions

The dimensionality of the Wisconsin Breast Cancer dataset, which is increasingly recognized as the "gold standard" for diagnosing malignant and benign tumors, can be significantly reduced without sacrificing predictive power. The attribute space dimension was reduced using the methods described in subsection 1.3 (first task).

The deep learning algorithms in WEKA demonstrate excellent performance in supervised and unsupervised learning, regardless of whether they are applied to full or reduced datasets. The WEKA experiment, which was planned for Task #2 in subsection 1.3, confirmed this finding.

In addition, the clustering of all datasets aligns well with the results obtained from classifications using deep learning algorithms.

Author Contributions: Conceptualization – **Gennady Chuiko**, **Denys Honcharov**; methodology – **Gennady Chuiko**, **Denys Honcharov**; simulation – **Gennady Chuiko**; validation – **Gennady Chuiko**; formal analysis – **Gennady Chuiko**, **Denys Honcharov**; Investigation – **Gennady Chuiko**; resources – **Gennady Chuiko**, **Denys Honcharov**; data curation – **Gennady Chuiko**; writing–original draft – **Gennady Chuiko**, **Denys Honcharov**; writing–review and editing – **Gennady Chuiko**; supervision – **Gennady Chuiko**; project administration – **Gennady Chuiko**;

Conflict of Interest

The authors declare that they have no conflict of interest concerning this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Financing

This study was conducted without financial support.

Data Availability

The data associated with this work are stored in the data repository (UCI Machine Learning Repository, 1995. DOI: 10.24432/C5DW2B) [7].

Use of Artificial Intelligence

The authors declare that they used artificial intelligence to test the text's spelling, grammar, style, and possible plagiarism using the AI-based Grammarly software. The authors have read, fixed, and polished the above version. Therefore, they are fully responsible for this text.

All the authors have read and agreed to the publication of the final version of this manuscript.

References

1. Orlova, N. M., Tonkovyd, O. B., Palamar, I. V., Klimas, L. A., Shkondin, S. V., & Tkach, V. S. Medykostatystychnyi analiz zakhvoriuvanosti, smertnosti ta svoiechasnosti vyiavlennia raku molochnoi zalozy v Ukraini [Medical and statistical analysis of incidence, mortality, and timeliness of breast cancer diagnosis in Ukraine]. *Visnyk Vinnytskoho natsionalnoho medychnoho universytetu – Rep. of Vinnytsia Nation. Med. Univ.*, 2024, vol. 28(1), pp. 113-120. DOI: 10.31393/reports-vnmedical-2024-28(1)-20. (In Ukrainian)

2. Zielonke, N, Kregting, L. M., Heijnsdijk, E. A. M., Veerus, P., Heinävaara, S., McKee, M., Kok, I. M. C. M., Koning, H. J., & Ravesteyn, N. T. The potential of breast cancer screening in Europe. *Int J Cancer*, 2021, vol. 148, iss. 2, pp. 406-418. DOI: 10.1002/ijc.33204.

3. Nusrat Mohi ud din, Rayees Ahmad Dar, Muzafar Rasool, & Assif Assad. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead, *Computers in Biology and Medicine*, 2022, vol. 149, article no. 106073, DOI: 10.1016/j.compbiomed.2022.106073.

4. Tolstoluzka, O., & Telezhenko, D., Development and Training of LSTM Models for Controlling Virtual Distributed Systems Using TensorFlow and Keras. *Radioelectronic and Computer Systems*, 2024, no. 1(109), pp. 27-37. DOI: 10.32620/reks.2024.3.02.

5. Zwitter, M., & Soklic, M. Breast Cancer. *UCI Machine Learning Repository*. Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia, 1998. DOI: 10.24432/C51P4M.

6. Chuiko, G. P., & Yaremchuk, O. M. Handling the Breast Cancer Recurrence Data for a More Reliable Forecast. *Kompiuterni systemy ta informatsiini tekhnolohii – Computer Systems and Information Technologies*, 2023, vol. (4), pp. 10-15. DOI: 10.31891/csit-2023-4-2.

7. Wolberg, W., Mangasarian, O., Street, N., & Street, W. Breast Cancer Wisconsin (Diagnostic). *UCI Machine Learning Repository*, 1995. DOI: 10.24432/C5DW2B.

8. BreakHis. *Breast Cancer Histopathological Database (BreakHis)*. Available at: https://www.kaggle.com/datasets/ambarish/breakhis (accessed 12.06.2024).

9. Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. A. Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, 2016, vol. 63, iss. 7, pp. 1455-1462. DOI: 10.1109/TBME.2015.2496264.

10. Nasser, M., & Yusof, U. K. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Diagnostics*, 2023, vol. 13, iss. 1, pp. 1-26. DOI: 10.3390/diagnostics13010161.

11. Nemade, V., Pathak, S., & Dubey, A. K. A Systematic Literature Review of Breast Cancer Diagnosis Using Machine Intelligence Techniques. *Archives of Computational Methods in Engineering*, 2022, vol. 29, pp. 4401-4430. DOI: 10.1007/s11831-022-09738-3.

12. Shashmi, K. *Curse of Dimensionality – A "Curse" to Machine Learning. Towards Data Science*. Available at: https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb (accessed 12.06.2024).

13. Chuiko, G. P., Honcharov, D. S., Dvornik, O. V., Krainyk, Ya. M., Darnapuk, Ye. O., & Yaremchuk, O. M. Attribute Selection, Outliers Impact Study, and Visualization within Breast Cancer Detection. *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, Lviv, Ukraine, 2023, pp. 1-5. DOI: 10.1109/ELIT61488.2023.10310922.

14. Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958, vol. 65, iss. 6, pp. 386-408. DOI: 10.1037/h0042519.

15. Freund, Y, & Schapire, R. E. Large Margin Classification Using the Perceptron Algorithm. *Mach Learn*, 1999, vol. 37, iss. 3, pp. 277–296. DOI: 10.1023/A:1007662407062.

16. Lang, S., Bravo-Marquez, F., Beckham, C., Hall, M., & Frank, E. WekaDeeplearning4j: A Deep Learning package for Weka based on Deeplearning4j. *Knowledge-Based Syst.*, 2019, vol. 178, pp. 48-50. DOI: 10.1016/j.knosys.2019.04.013.

17. Nova, D., & Estevez, P. A review of learning vector quantization classifiers. *Neural Computing and Applications*, 2014, vol. 25, pp. 511-524. DOI: 10.1007/s00521-013-1535-3.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

18. Wehrens, R., & Kruisselbrink, J. W. Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, 2018, vol. 87, pp. 1-18. DOI: 10.18637/JSS.V087.I07.

19. *Running an experiment using clusterers*. Available at: https://waikato.github.io/weka-wiki/experimenter/running_an_experiment_using_clusterers/ (accessed 12.06.2024).

20. Chicco, D., & Jurman, G. A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index. *J Biomed Inform.*, 2023, vol. 144, article no. 104426. DOI: 10.1016/j.jbi.2023.104426.

21. Xu, D., & Tian, Y. A. Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci.* 2015, vol. 2, pp. 165-193. DOI: 10.1007/s40745-015-0040-1.

22. *Calinski-Harabasz Index – Cluster Validity indices*. Available at: https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/ (accessed 12.06.2024).

23. Fowlkes, E. B., & Mallows, C. L. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 1983, vol. 78, iss. 383, pp. 553-569. DOI: 10.2307/2288117.

24. Ajayan, S. S. J. , Reddy, N. V. U., Devasenapati, S. B., & Rebelli, S. Analysis of COVID-19 CT Chest Image Classification using Dl4jMlp Classifier and Multilayer Perceptron in WEKA Environment. *Curr Med Imaging Former Curr Med Imaging Rev.*, 2023, vol. 20, pp. 1-7. DOI: 10.2174/1573405620666230417090246.

Received 17.07.2023, Accepted 18.11.2024

ЕФЕКТИВНІСТЬ АЛГОРИТМІВ СКОРОЧЕННЯ ВИМІРНОСТІ ТА ГЛИБОКОГО НАВЧАННЯ ЩОДО НАБОРУ ДІАГНОСТИЧНИХ ДАНИХ РАКУ МОЛОЧНОЇ ЖИВКИ

Г. П. Чуйко, Д. С. Гончаров

Рак молочної залози є серйозною загрозою, оскільки її найбільш часто діагностують та вона є однією з основних причин смертності серед жінок. Рання діагностика та своєчасне лікування мають вирішальне значення для збереження життя пацієнтові та зниження витрат на лікування. Різноманітні медичні методи візуалізації, такі як мамографія, комп'ютерна томографія, гістопатологія та ультразвук, є сучасними підходами до виявлення та класифікації раку молочної залози. Фахівці з машинного навчання віддають перевагу алгоритмам глибокого навчання для аналізу значних даних медичних зображень. Однак застосування діагностичних методів на основі глибокого навчання в клінічній практиці все ще обмежене, незважаючи на їх потенційну ефективність. Хоча методи глибокого навчання складні та непрозорі, їхня ефективність може допомогти збалансувати ці проблеми. **Предмети дослідження**. Алгоритми глибокого навчання, реалізовані в програмному забезпеченні WEKA, і їхня ефективність щодо Вінсконсінського набору даних про рак молочної залози. **Мета**. Значне зменшення розмірності набору даних без втрати прогностичної потужності. **Методи**. Комп'ютерні експерименти в середовищі WEKA забезпечують попередню обробку, контрольоване та неконтрольоване глибоке навчання повних і скорочених наборів даних з оцінкою їх ефективності. **Результати**. Потрійна послідовна фільтрація дозволила помітно скоротити розмірність вихідного набору даних: від 30 атрибутів до чотирьох. Несподівано всі три класифікатори глибокого навчання, реалізовані в WEKA (Dl4jMlp, Multilayer Perceptron і Voted Perceptron), показали статистично однакову продуктивність. Крім того, продуктивність була статистично однаковою для повних і скорочених наборів даних. Наприклад, відсоток правильно класифікованих екземплярів був у діапазоні (95,9-97,7) зі стандартним відхиленням менше 2,5 %. Два алгоритми кластеризації, які використовують нейрони (Self Organized Map, SOM, і Learning Vector Quantization, LVQ), також показали подібні результати. Два кластери в усіх наборах даних не розділені належним чином, але вони точно представляють обидва попередньо призначені класи з індексами Фаулкса–Меллоу (FMI) у діапазоні від 0,81 до 0,99. **Висновки**. Дослідження показує, що розмірність Вінсконсінського набору даних про рак молочної залози, який все більше стає «золотим стандартом» для діагностики злоякісних і доброякісних пухлин, може бути значно зменшена без втрати прогностичної потужності. Алгоритми глибокого навчання в WEKA забезпечують чудову продуктивність як для контрольованого, так і для неконтрольованого навчання, незалежно від того, чи йдеться про повні або скорочені набори даних.

Ключові слова: рак молочної залози; алгоритми глибокого навчання; WEKA; набір даних раку молочної залози Вісконсінського університета; діагностика злоякісних і доброякісних пухлин.

Чуйко Геннадій Петрович – д-р фіз-мат.. наук, проф., проф. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Гончаров Денис Сергійович – асп. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Gennady Chuiko – D.Sc. in Physics and Mathematics, Professor at the Computer Engineering Department, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine,
e-mail: henadiy.chuyko@chmnu.edu.ua, ORCID: 0000-0001-5590-9404.

Denys Honcharov – PhD Student of the Computer Engineering Department, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine, e-mail: honcharov.denys@chmnu.edu.ua, ORCID: 0009-0004-1200-6677.