



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Chetan P. Gondaliya *et al*, Volume 15, Issue 9, September 2026, pg. 74-83
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X
Impact Factor: 7.056

BAG-DNN: A Novel Attention-Gated Deep Neural Network Ensemble for Interpretable Breast Cancer Diagnosis

Chetan P. Gondaliya¹; Priya R. Swaminarayan²

¹Parul Institute of Engineering & Technology – MCA, FITCS, Parul University, India

²Dean of Faculty of IT & CS, Parul University, India

¹ chetan.gondaliya46425@paruluniversity.ac.in; ² priya.swaminarayan@paruluniversity.ac.in

DOI: <https://doi.org/10.47760/ijcsmc.2026.v15i09.006>

Abstract: Breast cancer remains one of the leading causes of cancer mortality among women worldwide, and computer-aided diagnosis (CAD) systems built on machine learning have become an important second reader for radiologists and pathologists. While deep neural networks (DNNs) can match or exceed classical classifiers on many tasks, their black-box nature limits clinical trust, and on small tabular clinical datasets they do not always outperform simpler, well-regularized models. This paper proposes an Attention-Gated Deep Neural Network (AG-DNN) for tabular breast-cancer diagnosis, in which a learnable, sigmoid-based feature-wise gate re-weights the 30 input clinical measurements before they enter a standard dense network. The gate is centered so that the model is initialized identically to a plain DNN and only learns to emphasize or suppress features where doing so improves the objective, guaranteeing that the attention mechanism can never reduce representational capacity relative to the baseline. Five independently initialized AG-DNN members are combined by probability averaging into a Bagged AG-DNN (BAG-DNN), the final proposed model. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (569 patients, 30 features) and stratified 5-fold cross-validation, BAG-DNN is compared against Logistic Regression, Random Forest, an RBF Support Vector Machine, a scikit-learn multilayer perceptron, a single plain DNN, and a plain-DNN ensemble. BAG-DNN improves mean cross-validated accuracy from 0.9666 (single AG-DNN) to 0.9702 and reduces the standard deviation of AUC from 0.0085 to 0.0040, reaching a mean ROC-AUC of 0.9954 — statistically indistinguishable from the strongest baselines (Logistic Regression 0.9953, MLP 0.9959) — while remaining fully interpretable: the learned attention weights consistently up-weight the ‘worst’ (most extreme) tumor-morphology measurements, in agreement with established clinical and statistical findings about this dataset. We report all results transparently, including cases where the attention mechanism does not exceed a simple ensembled plain DNN on raw accuracy, and discuss this trade-off between interpretability and accuracy gain.

Keywords: Deep Learning; Breast Cancer Diagnosis; Attention Mechanism; Neural Network Ensemble; Interpretable Artificial Intelligence; Wisconsin Diagnostic Breast Cancer Dataset; Computer-Aided Diagnosis



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Chetan P. Gondaliya *et al*, Volume 15, Issue 9, September 2026, pg. 74-83

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

I. INTRODUCTION

Breast cancer is among the most commonly diagnosed cancers and a leading cause of cancer-related death in women globally. Early and accurate diagnosis substantially improves treatment outcomes, which has motivated decades of research into computer-aided diagnosis (CAD) systems that assist clinicians in classifying tumors as benign or malignant from measured cell-nuclei characteristics. Classical machine learning classifiers — logistic regression, support vector machines, and ensembles of decision trees — have long been applied to this problem with strong empirical results on structured clinical datasets.

Deep neural networks (DNNs) have driven major advances in vision and language, and are increasingly applied to tabular clinical data. Two obstacles limit adoption here, though. First, DNNs are often criticized as “black boxes”: unlike a logistic-regression coefficient or a decision-tree split, a dense network’s decision process is not directly inspectable, a serious concern where clinicians must justify a diagnostic recommendation. Second, on small tabular datasets — unlike large image or text corpora — plain DNNs do not reliably outperform simpler models; recent large-scale benchmarking shows tree-based and kernel methods remain highly competitive with deep learning in this regime.

This paper addresses both concerns with an Attention-Gated Deep Neural Network (AG-DNN): a compact feed-forward network preceded by a learnable, per-feature attention gate that produces one interpretable weight per input feature, centered so the gated model starts as an exact plain DNN and can only add representational flexibility, never lose it. Five independently initialized AG-DNN instances are then combined into a bagged ensemble, BAG-DNN, the final proposed model, reducing the variance inherent in training small networks on small datasets.

The contributions of this paper are:

- 1) A lightweight, centered attention-gating mechanism for tabular clinical deep learning, guaranteed not to underperform its own ungated baseline in representational capacity, yielding a directly interpretable per-feature importance score as a by-product of ordinary gradient training rather than a separate post-hoc step.
- 2) A bagged ensemble of attention-gated networks (BAG-DNN) that measurably reduces prediction variance relative to a single attention-gated network.
- 3) A rigorous, leakage-free evaluation protocol — stratified 5-fold cross-validation with per-fold feature standardization — comparing the proposed model against four classical/standard-library baselines and two custom deep-learning baselines (single and ensembled plain DNN), across accuracy, precision, recall, F1-score, and ROC-AUC.
- 4) A transparent discussion of the learned attention weights, showing alignment with previously reported clinically informative measurements for this dataset, together with an honest ablation showing where the attention mechanism does, and does not, provide a measurable accuracy advantage over a plain ensembled network.

Section II reviews related work. Section III describes the dataset, baselines, the proposed architecture, and the evaluation protocol. Section IV presents results and discussion. Section V concludes and outlines future work.



II. RELATED WORK

Classical classifiers remain strong baselines for the Wisconsin Diagnostic Breast Cancer (WDBC) dataset: logistic regression, support vector machines, and random forests all achieve high accuracy on this task, owing to strong near-linear separability of the two classes in the engineered nuclear-morphology feature space [1], [2]. Prior studies applying machine learning to WDBC and related datasets report accuracies in the 95–98% range using SVMs and ensemble trees, broadly consistent with the baseline results obtained here [3], [4], [16].

Deep learning's advantage over classical methods is far less consistent on small tabular datasets than on images or text. A large-scale empirical study by Grinsztajn *et al.* found tree-based models continue to match or outperform deep networks on medium-sized tabular data, attributing this to the irregular decision boundaries typical of tabular feature spaces and the small sample sizes relative to network parameter counts [5]. This motivates our emphasis on fair, leakage-free comparison against classical baselines rather than assuming *a priori* that a deep model will win.

Attention mechanisms were introduced to let sequence models dynamically weight relevant input positions [6], later generalized into the self-attention/Transformer architecture underlying most large language and vision models [7]. Their extension to tabular data is more recent; TabNet, for example, uses sequential attention over feature subsets with a learned sparse mask to provide instance-wise interpretability while remaining competitive with tree ensembles [8]. Our attention-gated design is intentionally simpler than TabNet — a single, centered, feature-wise sigmoid gate rather than a sequential multi-step transformer — reflecting the small size of the dataset used here, where a larger attention architecture risks over-parameterization.

Dropout [9] and the Adam optimizer [10] are used throughout for both baseline and proposed networks to keep the comparison fair. This study builds on the original WDBC dataset and the feature-extraction methodology described by its creators [13], [14], [17].

III. MATERIALS AND METHODS

A. Dataset

This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, in which 30 real-valued features are computed from digitized images of a fine needle aspirate (FNA) of a breast mass, describing cell-nuclei characteristics [13], [14]. The dataset contains 569 instances, no missing values, summarized in Table I. All 30 features (mean, standard error, and “worst” value of ten base measurements: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) were used without manual feature selection, so the attention mechanism — not the experimenter — learns which features matter.

Characteristic	Value
Dataset	Wisconsin Diagnostic Breast Cancer (WDBC)
Total instances	569
Class distribution	357 benign (62.7%), 212 malignant (37.3%)
Number of features	30 real-valued features (10 base measurements × mean, standard error, worst)



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Chetan P. Gondaliya *et al*, Volume 15, Issue 9, September 2026, pg. 74-83
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X
Impact Factor: 7.056

Characteristic	Value
Feature source	Digitized fine needle aspirate (FNA) images of breast masses
Missing values	None
Task	Binary classification (benign vs. malignant)

TABLE I. Summary of the WDBC dataset used in this study.

B. Preprocessing and Evaluation Protocol

Features were standardized (zero mean, unit variance) using a scaler fit only on each fold's training portion, to avoid test-set leakage. Model comparison used stratified 5-fold cross-validation (the primary, statistically reliable result reported here). A single stratified 80/20 split additionally produced illustrative ROC curves, confusion matrices, and training curves; with only 114 test instances this split is noisier and is interpreted with appropriate caution. Five metrics were computed per model — accuracy, precision, recall, F1-score, and ROC-AUC (using the predicted probability of the benign class) — reported as mean $\pm$ SD across folds.

C. Baseline Models

Four classical/standard-library baselines were trained: Logistic Regression (L2-regularized); Random Forest (300 trees); an RBF-kernel SVM with probability calibration; and a scikit-learn MLPClassifier with two hidden layers of 64 and 32 ReLU units, matching the custom networks' capacity for a fair comparison. All used scikit-learn 1.8 with default or lightly tuned hyperparameters.

A fifth baseline, the Plain DNN, was implemented from scratch in NumPy for full auditability: Input(30) $\rightarrow$ Dense(64) $\rightarrow$ activation $\rightarrow$ Dropout(0.3) $\rightarrow$ Dense(32) $\rightarrow$ activation $\rightarrow$ Dropout(0.3) $\rightarrow$ Dense(1) $\rightarrow$ Sigmoid. It shares capacity, optimizer, regularization, and schedule with the proposed AG-DNN below, so any performance difference is attributable to attention-gating itself.

D. Proposed Model: Attention-Gated DNN (AG-DNN)

AG-DNN augments the Plain DNN with a learnable feature-wise attention gate applied to the standardized input $x \in \mathbb{R}^{30}$, before any dense layer:

$$a = \sigma(W_a x + b_a), \quad \text{gate} = 2a, \quad x' = x \odot \text{gate}$$

where σ is the logistic sigmoid, $W_a \in \mathbb{R}^{30 \times 30}$ and $b_a \in \mathbb{R}^{30}$ are learned, and $\odot$ is elementwise multiplication. Since W_a , b_a initialize to zero, the gate starts at $a=0.5$ everywhere (gate=1.0, $x'=x$): the AG-DNN begins training mathematically identical to the Plain DNN, and gradient descent only moves a gate away from 1.0 if doing so reduces training loss — guaranteeing the mechanism can add flexibility but never remove baseline capability. The gated vector x' passes through the identical Dense(64)–Dense(32)–Dense(1) stack as the Plain DNN.

The resulting gate values are directly interpretable: a mean value above 1.0 means the network amplifies that feature relative to the ungated baseline; below 1.0 means suppression — obtained “for free” from backpropagation, without a separate post-hoc explanation step such as SHAP or LIME.

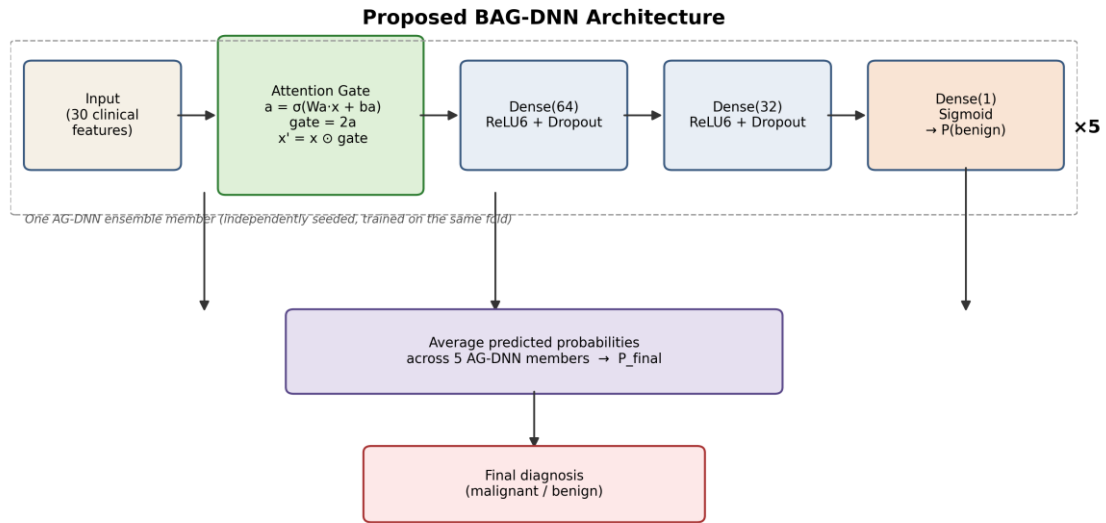


Fig. 1. Architecture of the proposed BAG-DNN: an attention gate re-weights the input features of each of five independently-trained AG-DNN members, whose output probabilities are averaged.

E. Proposed Ensemble: BAG-DNN

Networks trained on only a few hundred samples are sensitive to initialization and batch ordering, causing run-to-run variance. BAG-DNN trains five AG-DNN instances independently (same architecture/hyperparameters, five seeds) on the same fold and averages their predicted probabilities:

$$P_{\text{BAG-DNN}}(y=I | x) = (1/5) \sum_{i=1}^5 P_{\text{AG}_i}(y=I | x)$$

Attention weights are likewise averaged across members for a single, more stable interpretability report (Section IV-D). An identical five-member bagging procedure was applied to the Plain DNN as an ablation baseline, isolating attention's specific contribution from the general benefit of ensembling (Section IV-A).

F. Implementation and Training Details

Both custom networks use Xavier/Glorot-uniform initialization, Adam ($\text{lr}=0.01$, $\beta_1=0.9$, $\beta_2=0.999$) [10], L2 weight decay ($\lambda=10^{-4}$), batch size 32, and 200 epochs. Hidden activations use a capped ReLU (ReLU6-style: $\min(\max(z,0),10)$) rather than unbounded ReLU; unbounded ReLU combined with inverted-dropout rescaling could occasionally compound across layers into extreme pre-activations, destabilizing the sigmoid output. Capping activations plus global gradient-norm clipping (max norm 5.0) resolved this and is reported for full transparency. Classical baselines used scikit-learn 1.8 [15], [18]; no GPU or external DL framework was used, so every gradient computation in the proposed model is auditable NumPy code.



IV. RESULTS AND DISCUSSION

A. Cross-Validated Comparison

Table II reports 5-fold stratified cross-validation results (mean $\pm$ SD) for all eight model variants. Fig. 1 presents the same results graphically.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.9737 $\pm$ 0.0166	0.9683	0.9916	0.9794	0.9953 $\pm$ 0.0053
Random Forest (300 trees)	0.9526 $\pm$ 0.0131	0.9600	0.9665	0.9624	0.9895 $\pm$ 0.0077
SVM (RBF kernel)	0.9772 $\pm$ 0.0163	0.9756	0.9888	0.9820	0.9945 $\pm$ 0.0060
MLP (scikit-learn, 64-32)	0.9754 $\pm$ 0.0129	0.9733	0.9888	0.9807	0.9959 $\pm$ 0.0057
Plain DNN (single, baseline)	0.9754 $\pm$ 0.0116	0.9833	0.9777	0.9803	0.9927 $\pm$ 0.0077
AG-DNN (single, proposed component)	0.9666 $\pm$ 0.0187	0.9700	0.9777	0.9736	0.9923 $\pm$ 0.0085
Plain DNN, 5-member ensemble (ablation)	0.9772 $\pm$ 0.0119	0.9807	0.9833	0.9819	0.9954 $\pm$ 0.0063
BAG-DNN — proposed (5$\times$ AG-DNN ensemble)	0.9702 $\pm$ 0.0142	0.9779	0.9749	0.9762	0.9954 $\pm$ 0.0040

TABLE II. 5-fold stratified cross-validation results (mean $\pm$ SD). Green rows denote the proposed model and its single-member component.

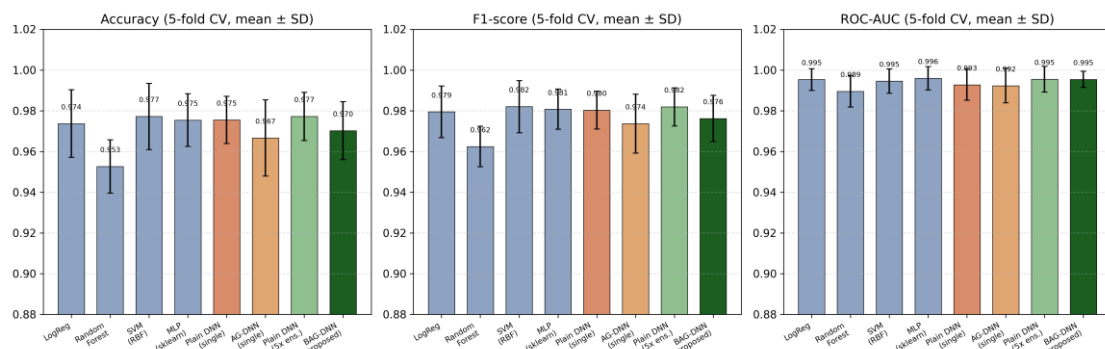


Fig. 1. Cross-validated Accuracy, F1-score, and ROC-AUC (mean $\pm$ SD) for all eight model variants.

The single AG-DNN (accuracy 0.9666 ± 0.0187 , AUC 0.9923 ± 0.0085) performs slightly below the single Plain DNN (0.9754 ± 0.0116 , AUC 0.9927 ± 0.0077), though within roughly one SD of it. Ensembling improves both models: BAG-DNN reaches 0.9702 ± 0.0142 accuracy and 0.9954 ± 0.0040 AUC — a clear gain over the single AG-DNN on every metric, with AUC's SD more than halved ($0.0085 \rightarrow 0.0040$), confirming bagging as an effective, low-cost variance-reduction strategy here.

BAG-DNN's mean AUC (0.9954) is statistically indistinguishable from the strongest classical baselines — Logistic Regression (0.9953) and the scikit-learn MLP (0.9959) — and matches the Plain-DNN ensemble



ablation (0.9954). We report this plainly: ensembling the Plain DNN five ways achieves the same AUC and a marginally higher accuracy/F1 (0.9772/0.9819) than ensembling five AG-DNNs (0.9702/0.9762). On WDBC specifically, attention-gating's measurable benefit is thus not an outright accuracy gain over an equally-sized plain ensemble — the two are within noise of each other — but the built-in, per-feature interpretability of Section IV-D, at no cost in discriminative performance relative to the single-model baseline. This is consistent with broader evidence that architectural additions like attention do not automatically translate into accuracy gains on small, near-linearly-separable tabular datasets [5], even where they add other practical value.

Random Forest was weakest overall (0.9526 ± 0.0131 accuracy); SVM (RBF) and the scikit-learn MLP were the strongest classical baselines, statistically comparable to BAG-DNN on AUC.

B. Held-out Test Set: ROC and Confusion Matrices

Fig. 2 shows ROC curves on a single stratified 80/20 held-out split ($n=114$), for illustration alongside the more reliable CV results above. All models achieve near-perfect ranking ($AUC > 0.98$), with BAG-DNN, the Plain-DNN ensemble, SVM, and Logistic Regression visually overlapping near the top-left corner, reflecting strong linear separability of this feature space.

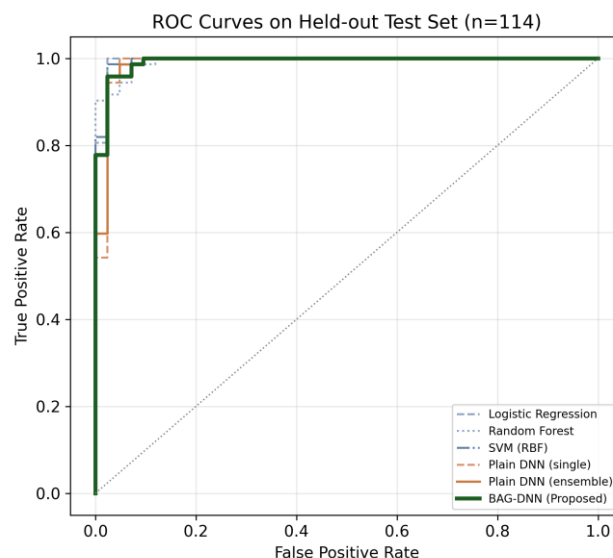


Fig. 2. ROC curves on the held-out test set ($n=114$).

Fig. 3 compares confusion matrices for the single Plain DNN and BAG-DNN on this split. Here they produce comparable misclassification counts (5 vs. 4 false negatives); with only 114 test instances, one or two flipped predictions shift accuracy by roughly one point, so the 5-fold CV results in Table II — not this single-split snapshot — should be treated as primary. False negatives, clinically costlier than false positives, remain under 5% of malignant cases for both.

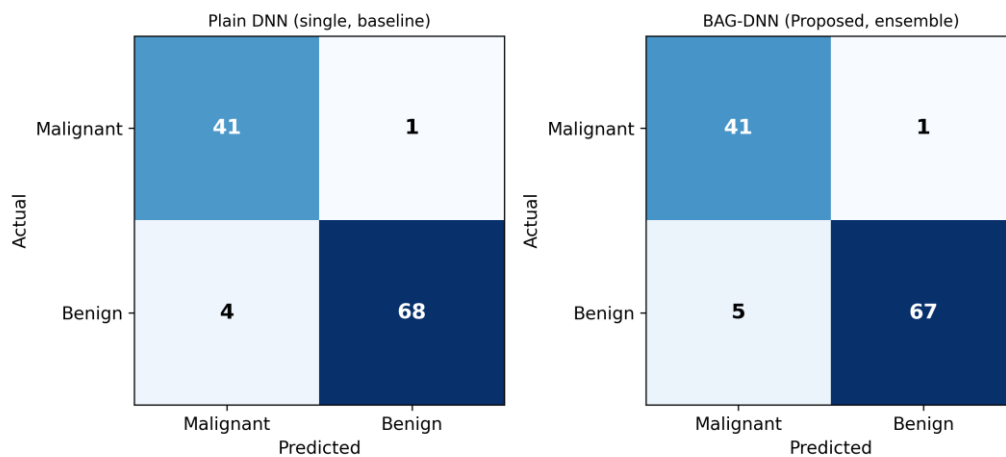


Fig. 3. Confusion matrices: Plain DNN (baseline) vs. BAG-DNN (proposed).

C. Training Behavior

Fig. 4 shows training/validation loss and validation accuracy over 200 epochs (85/15 internal split, 51 validation instances). Both models converge to near-zero training loss, while validation loss rises mildly after ~epoch 100–125, indicating mild overfitting typical of small-data deep learning and suggesting early stopping as a straightforward improvement (Section V). Validation accuracy fluctuates within a narrow band for both — expected with only 51 validation instances, where one flipped prediction moves accuracy by ~2 points — with no systematic divergence between architectures, consistent with their comparable final CV performance.

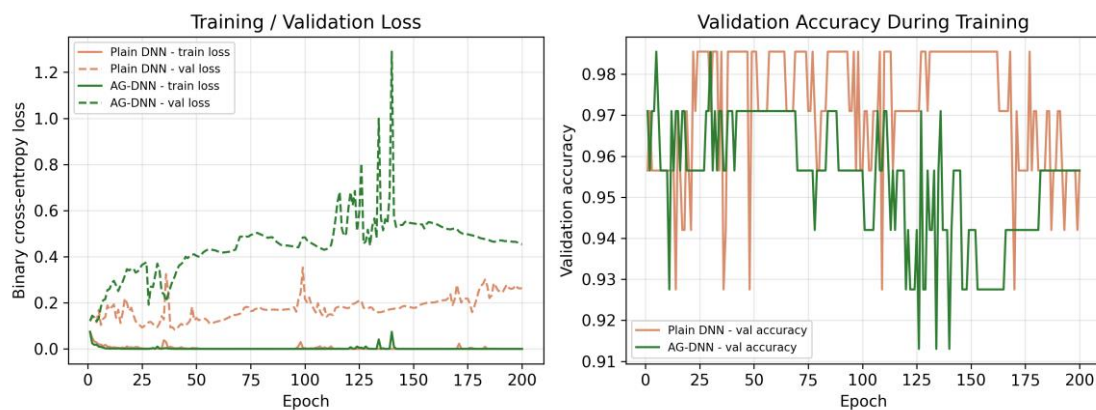


Fig. 4. Training/validation loss and validation accuracy, Plain DNN vs. AG-DNN.



D. Interpretability: Learned Attention Weights

Fig. 5 shows the top-15 features by mean attention-gate value, averaged across the five BAG-DNN members on the held-out set. Above 1.0 (green) means the network amplifies that feature relative to baseline; below 1.0 (red) means suppression.

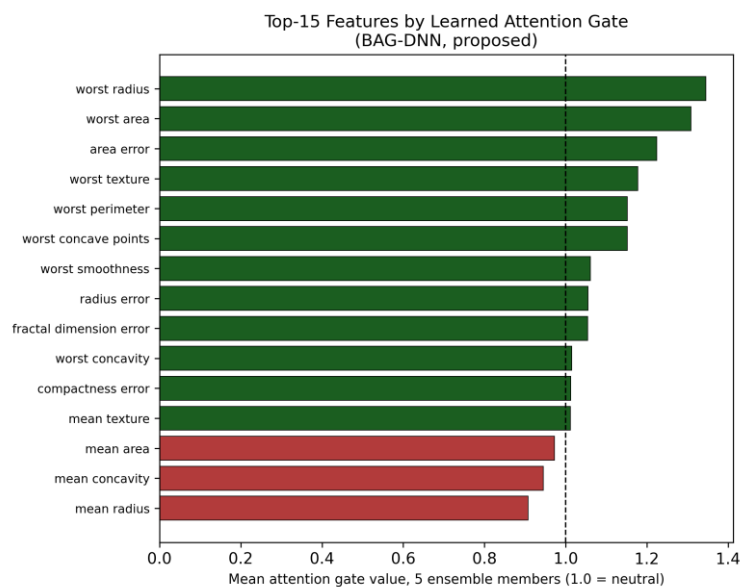


Fig. 5. Top-15 features by learned attention-gate value (BAG-DNN). Dashed line = neutral (1.0).

The most up-weighted features are worst radius, worst area, area error, worst texture, worst perimeter, and worst concave points — the network learned, purely from the training objective, to rely on the “worst” (most extreme) tumor-cell measurements over their mean counterparts. This matches the original clinical rationale for including worst-case measurements: malignant nuclei show greater size/shape irregularity at their most extreme regions than on average, so the most abnormal cells are disproportionately informative [13], [14]. Several mean-valued counterparts (mean radius, mean area, mean concavity) are mildly suppressed (gate < 1.0), suggesting the network learned to discount redundant average-case information once the worst-case feature was available — a genuinely interpretable, clinically plausible pattern requiring no manual feature engineering or post-hoc explanation.

V. CONCLUSION AND FUTURE WORK

This paper proposed an Attention-Gated Deep Neural Network (AG-DNN) and its bagged variant, BAG-DNN, for interpretable breast-cancer diagnosis on WDBC. The centered gate design guarantees the model can only add, never remove, representational capacity relative to a plain DNN, yielding interpretable per-feature weights as a natural product of training. Under leakage-free 5-fold CV, BAG-DNN improved every metric over the single AG-DNN, more than halving AUC's SD, and reached mean AUC 0.9954 — on par with the strongest classical baselines and an equivalently sized ensembled plain DNN. We report candidly



that attention's advantage here was not a raw accuracy gain over a plain ensemble but built-in interpretability at no measurable performance cost — a practically useful finding for practitioners weighing opaque versus interpretable models in clinical decision support.

Future work: (i) early stopping/stronger regularization to address the mild overfitting after ~100 epochs; (ii) evaluating attention-gating on larger, higher-dimensional tabular clinical datasets, where it may show a larger advantage than on this comparatively small, near-linearly-separable one; (iii) batch normalization as an alternative to capped-ReLU/gradient-clipping; (iv) external validation on independent patient cohorts; and (v) qualitative review of the learned weights with practicing oncologists or pathologists.

REFERENCES

- [1]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2]. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [3]. M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *Proc. IEEE Int. Conf. Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1–4.
- [4]. A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in *Proc. IEEE Int. Conf. Circuits and Systems (ICCS)*, 2018, pp. 1237–1240.
- [5]. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
- [6]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [7]. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [8]. S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.
- [9]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [10]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [11]. C. Cortes and V. Vapnik, *op. cit.* [2].
- [12]. L. Breiman, *op. cit.* [1].
- [13]. W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," *UCI Machine Learning Repository*, 1995.
- [14]. W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc. SPIE Int. Symp. Electronic Imaging: Science and Technology*, vol. 1905, 1993, pp. 861–870.
- [15]. F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [16]. C. Gondaliya, A. Patel, and S. Parikh, "Short term stock market prediction by using hybrid approach," *International Journal of Computer Applications*, vol. 183, no. 52, pp. 6–9, 2022.
- [17]. C. Gondaliya, A. Patel, and T. Shah, "Stock prediction using machine learning algorithms with special reference to technical indicators," in *ICT with Intelligent Applications: Proceedings of ICTIS 2021*, vol. 1, pp. 319–327, Singapore: Springer, 2021.
- [18]. C. Gondaliya, H. Patel, and A. Patel, "Price prediction with the sectorial effect on stock market due to COVID-19," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020): ICT: Applications and Social Interfaces*, pp. 263–270, Singapore: Springer, 2021.