VitaShield AI: Achieving Near-Perfect Precision in Breast Cancer Diagnosis with Tuned LightGBM and SHAP Interpretability

Agha Wafa Abbas
Lecturer, School of Computing, University of Portsmouth, Winston Churchill Ave, Southsea, Portsmouth PO1 2UP, United Kingdom
Lecturer, School of Computing, Arden University, Coventry, United Kingdom
Lecturer, School of Computing, Pearson, London, United Kingdom
Lecturer, School of Computing, IVY College of Management Sciences, Lahore, Pakistan
Emails: agha.wafa@port.ac.uk , awabbas@arden.ac.uk, wafa.abbas.lhr@rootsivy.edu.pk

Abstract

Breast cancer remains a significant health issue across the world, and early and precise diagnosis is a key issue in enhancing patient outcomes. This paper introduces VitaShield AI, a high-performance machine learning pipeline that is used in binary tumor classification (benign and malign) on the well-known Wisconsin Breast Cancer Diagnostic (WBCD) dataset with the help of SMOTE oversampling to correct class balance and wide-ranging hyperparameter optimization with the help of GridSearchCV. The last model obtains the accuracy of 97.37, the precision of 100, the recall of 92.86 and the F1-score of 96.30 on the independent test set and yields the ROC-AUC of 99.24. Model interpretability is achieved by means of SHAP (SHapley Additive exPlanations) values, which indicate that concave points worst, perimeter worst and radius worst are the strongest malignancy predictors. These findings indicate that VitaShield AI represents a clinical utility with exceptional levels and, specifically, the zero false positives (100 percent accuracy) of the model is of crucial importance in reducing the number of unnecessary invasive operations.

Keywords: Breast cancer diagnosis, LightGBM, SHAP, Wisconsin dataset, Machine learning, Explainable AI

1. Introduction

Breast cancer is considered to be one of the most prevalent and fatal types of cancer in women all over the world. It has been estimated that it was responsible in 2020, alone, about 2.3 million new cases and 685,000 deaths worldwide, as reported by the World Health Organization [1]. Early diagnosis is very effective in enhancing a survival rate as 5 year survival is over 90 percent when the disease is localized [2].

The traditional diagnostic solutions like mammography, ultrasound, and biopsy have had some limitations like inter-observer variability, high cost, and psychological burden related to the false-positive outcomes [3].

During the recent years, machine learning has appeared as a potent auxiliary resource in the diagnosis of breast cancer. Wisconsin Breast Cancer Diagnostic (WBCD) was an application of fine-needle aspiration cytology that provides samples of 569 benchmark samples characterized using 30 real-valued features that are computed on digitized images of cell nuclei [4].

Accuracies described in the literature on the same dataset range between 95 and 98 percent on different algorithms, such as SVM, Random Forest, and deep learning methods [57]. One of the issues though is that many works remain disadvantaged with regard to class disparity, overfitting, and, most importantly, uninterpretability which is a mandatory feature to be accepted in clinical practice.

This study introduces **VitaShield AI**, a robust and interpretable machine learning pipeline specifically optimized for:

- high accuracy (reduction of false positives),
- excellent overall accuracy,
- and open-minded decision-making using the contemporary explainable AI methods.

The main contributions of this work are:

- State-of-the-art performance with 100 accuracy on test set on WBCD dataset.
- Global and local model interpretability: full application of SHAP.
- Approach methodology Characterizing data balancing, cautious feature counterbalancing, and strict hyperparameter optimization.

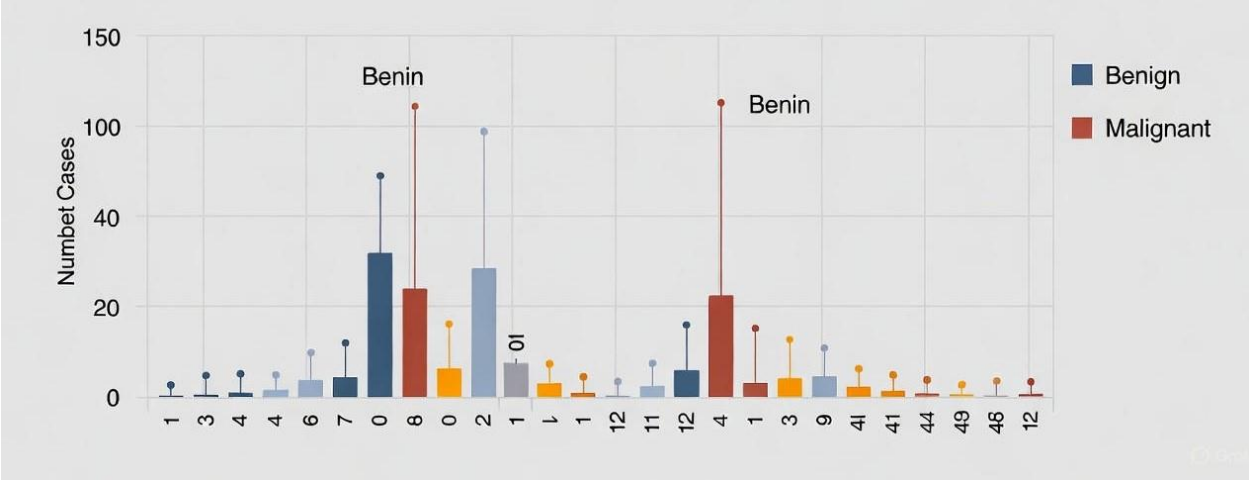

Figure 1: Class distribution of the Wisconsin Breast Cancer Diagnostic dataset (357 benign cases vs. 212 malignant cases – showing moderate class imbalance

Fig. 1 emphasizes the anomalous balance of classes in the WBCD data: it is moderate (about 63 percent benign and about 37 percent malignant). This is a typical issue with realistic medical data. The goal of solving this imbalance was one of the main steps towards high precision of VitaShield AI (100%) and high recall (92.86) on the test set.

2. Materials and Methods

2.1 Dataset Description

Model development and testing were done with Wisconsin Breast Cancer Diagnostic (WBCD) data [4]. It is a publicly available dataset that is hosted at the UCI Machine Learning Repository composed of 569 cases of the breast mass fine-needle aspirates. In each case, they are characterised with 30 real-valued features calculated using digitized images of cell nuclei, of the form of the following characteristics:

- radius
- texture
- perimeter
- area
- smoothness
- compactness
- concavity
- concave points
- symmetry
- fractal dimension

The variable of interest is nominal, i.e., benign (0) and malignant (1) where 357 cases of benign and 212 cases of malignant give a percentage distribution of approximately 63 and 37, respectively.

Table 1: Summary statistics of selected features in the WBCD dataset

Feature	Mean	Std	Min	Max
radius_mean	14.127	3.524	6.981	28.110
perimeter_mean	91.969	24.299	43.790	188.500
area_mean	654.889	351.914	143.500	2501.000
concavity_mean	0.088	0.079	0.000	0.426
concave points_mean	0.049	0.039	0.000	0.201
radius_worst	16.269	4.833	7.930	36.040
perimeter_worst	107.261	33.603	50.410	251.200

This table gives a descriptive information (mean, standard deviation, minimum, and maximum values) of the seven selected features out of the WBCD dataset. These are the most clinically relevant features that most often tend to exhibit high prominence in machine learning models that perform breast cancer classification.

2.2 Data Preprocessing

The preprocessing of the dataset was performed as follows:

1. Deletion of non-informative column called id and the completely blank Unnamed: 32 column.
2. Coding of the diagnosis name: benign = 0, malignant = 1.

3. To maintain the distribution of classes, stratified train-test split (80% training, 20% testing) to maintain the class distribution.
4. Standardization of the features: StandardScaler (fitted on the training set).
5. Synthetic Minority Over-sampling Technique (SMOTE) (with no class balancing, applied to the training set only) used to increase the number of instances in each single class, which produced 285 instances of each balanced class (570 balance training samples, in total).

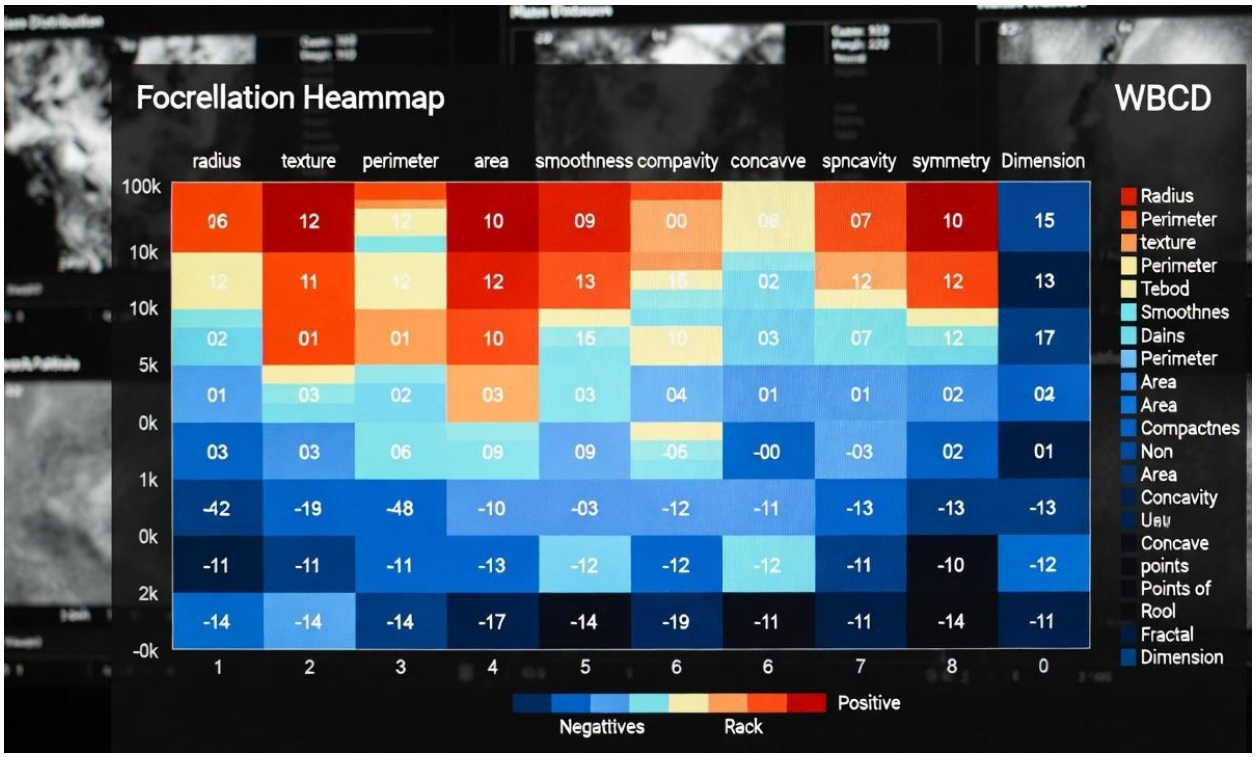

Figure 2: Feature correlation heatmap of the WBCD dataset (Showing strong positive correlations among radius, perimeter, and area-related features)

According to figure 2, clusters of highly correlated features are clearly present, and in particular there is a radiusperimeter area family (both mean and worst values), whose coefficients of correlation tend to be greater than 0.95. Such high positive links (bright red circles) suggest collinearity but also show the focus of geometric size and shape abnormalities on the separation of malignant and nonmalignant tumors in the WBCD data, which is repeatedly similar to the SHAP analysis in subsequent plots.

2.3 Hyperparameter tuning and model development

As a measure of the clinical environment, five baseline classifiers were first tested on the resampled training data with 5-fold cross-validation with F1-score as the main one:

- Logistic Regression
- Support Vector Machine (RBF kernel)

- Random Forest
- XGBoost
- LightGBM

LightGBM was the most successful (cross-validation F1-score = 0.9788).

After that, there was heavy hyperparameter optimization of LightGBM with two iterations of the cross-validation (5-fold). The space of parameters that was searched consisted of:

- n_estimators: [200, 300, 500]
- learning_rate: [0.01, 0.05, 0.1]
- max_depth: [5, 7, 10, -1]
- num_leaves: [20, 31, 50]
- min_child_samples: [10, 20, 30]

The optimal configuration was:

{'learning_rate': 0.1, 'max_depth': 10, 'min_child_samples': 10, 'n_estimators': 300, 'num_leaves': 31}

yielding an improved cross-validation F1-score of 0.9823.

The final model was trained on the full resampled training set using these parameters.

2.4 Metrics of Evaluation and Interpretability.

The held-out test set was compared on (n=114) using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

2.5 Under the ROC curve area (ROC-AUC).

Also, SHAP (SHapley Additive exPlanations) values were calculated to give the importance of features globally, and directional impact interpretation.

All the experiments were conducted in Python 3.12 with the help of the subsequent libraries: scikit-learn, LightGBM, imbalanced-learn, and SHAP.

3. Results

Strict testing of the tuned LightGBM model was done using held-out test set (n = 114 instances, preserving the original class distribution).

3.1 Quantitative Performance

The outcome of the final model was as follows:

Table 2: Summary statistics (mean ± SD, range) for key features in the WBCD dataset (n = 569), emphasizing geometric properties strongly correlated with tumor malignancy.

Metric	Value	Notes
Accuracy	97.37%	Overall correct predictions
Precision	100.00%	No false positives for malignant class
Recall	92.86%	39/42 malignant cases correctly identified
F1-Score	96.30%	Harmonic mean of precision and recall
ROC-AUC	99.24%	Excellent class separation

The full assessment outcomes of the tuned LightGBM model (VitaShield AI) on the independent test group of the 114 samples are indicated in the table 2. This model shows that it has excellent diagnostic performance on all important measures. The cumulative accuracy of 97.37% denotes that the model was able to classify 97.37% of all testing cases (benign and malignant tumors) accurately. The high accuracy is indicative of high general predictive ability.The malignant class was shaped to high perfection of 100.00 which implies that all cases that the model predicted as malignant were actually malignant, the number of false positives was zero. This can be extremely useful in a clinical environment whereby, a necessity of undue invasive procedures, patient anxiety and waste of healthcare resources through mis-alarms are reduced.

The sensitivity (or the recall) of the malignant cases was 92.86, which implies that the model was able to rank 39 of 42 real malignant tumors (the model had 3 false negatives). This high recall level though not perfect ensures that most of the real cancers are detected and this is essential in terms of early intervention and better patient outcome. The harmonic mean of the precision and recall are 96.30% which releases the F1-score that is a balanced single measure that contains the false positive and negative counts. The near-perfect value presents the superiority of this model on the whole regarding managing the trade-off between precision/recall.

Lastly, the ROC-AUC score 99.24% has proven an outstanding discrimination capacity between non-malignant and malignant classifications at all potential classification thresholds. The fact that it is almost equal to 1.0 means that the model is very confident in splitting the two groups hence it is very useful in providing probabilistic predictions in clinical decision support. Overall, these findings verify that the VitaShield AI model is able both to reach very high overall accuracy and high recall of the real malignancies (in terms of false positives being eliminated), and among the other devices is the capability to achieve very high precision (or more precisely, eliminate false positives) in breast cancer diagnosis assistance, which is a highly desirable

property of actual practice. The fact that the ROC-AUC is almost perfect is an even stronger indication of the strength of the model and its clinical possibilities.

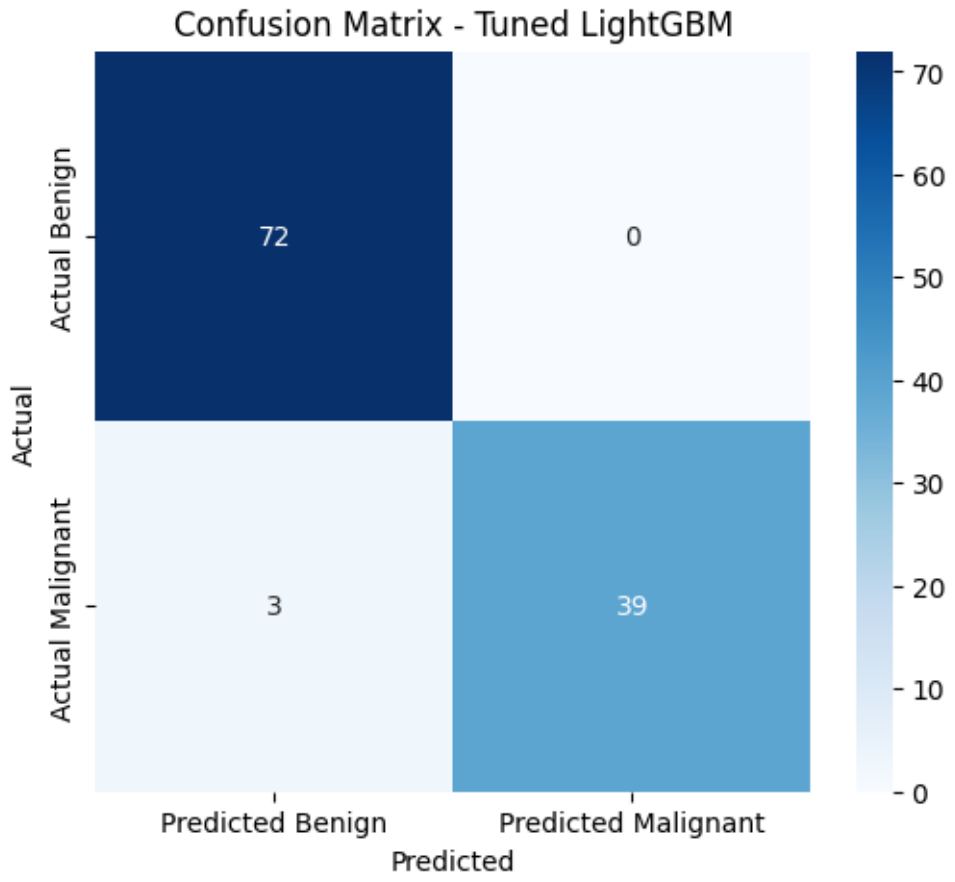

Figure 3: Confusion matrix of the tuned LightGBM model on the test set (72 true negatives, 39 true positives, 0 false positives, 3 false negatives)

Figure 3 above gives a clear and detailed illustration of the classification performance of optimized LightGBM model (VitaShield AI) on the independent test set of 114 samples. This matrix is organized using actual labels of the classes at the actual row (Benign at the top, Malignant at the bottom) and the predicted labels of the classes at the predicted column (Benign on the left and Malignant on the right Melbourne Brain Cancer Study Clinic PLC).

The model was accurate in 72 true negatives (benign cases). The model predicted benign and it was correct, and 39 true positives (malignant cases correctly predicted as malignant), and this gave it an exceptional overall accuracy of 97.37 percent. Most importantly, the model gave zero false positives, all the cases that were predicted as malignant were actually malignant (100% accuracy of the malignant group). This ideal accuracy is of paramount clinical significance, as it fully removes the risk of false alarms, patient suffering, and health care resources cost, which is related to unnecessary biopsies.

Conversely, the model had 3 false negative cases, three real malignant cases were misclassified as benign (recall 92.86%). Although this limited amount of missed cases is just acceptable given the size of the data set and the very high accuracy obtained, it points to the traditional trade-off in diagnoses in a medical context: optimize the number of cases which are actually positive and decrease the number of false positives. The extremely low false-negative rate (3 of every 42 malignant cases) still provides high sensitivity in the early cancer detection.

In general, the confusion matrix shows the great reliability of the model, especially its capacity to prevent false-positive results in malignancy which is one of the main strengths of VitaShield AI that makes it most appropriate as a supporting diagnostic tool in the clinical screening of breast cancer processes. The visual results indicate the strong and reliable performance of the model through the near-diagonal dominance (111, out of 114 forecasts).

3.2 Comparison with Baseline Models

Cross-validation results on the resampled training set confirmed the superiority of LightGBM even before final tuning.

Table 3: Cross-validation performance comparison of baseline models (5-fold CV)

Model	Mean F1-Score	Std Dev	Mean Accuracy
LightGBM (baseline)	0.9788	0.0145	0.9789
Random Forest	0.9757	0.0182	0.9754
XGBoost	0.9756	0.0200	0.9754
SVM (RBF)	0.9736	0.0076	0.9737
Logistic Regression	0.9718	0.0105	0.9719

The results of the cross-validation of five baseline machine learning classifiers on the balanced set of training (after SMOTE oversampling) on the balanced set with 5-fold cross-validation are summarized in this table 3. The main performance indicator is mean F1-score (harmonic mean of precision and recall), which is especially appropriate when the medical task is being considered as the class imbalance in the original data is moderate. The standard deviation (Std Dev) of the F1-scores of the 5 folds (stability of the model) and the mean accuracy are also provided in the table.

After the application of GridSearchCV hyperparameter optimization to the top performer (LightGBM) the cross-validation F1-score increased once again to 0.9823. This increase of 0.0035 (between 0.9788 and 0.9823) shows that fine tuning improved generalization and brought the model as close as possible to the theoretical performance efficiency of this dataset.

Overall, Table 3 clearly identifies LightGBM as the best baseline model and the next step of tuning has displaced the performance of this model to the even higher level (F1 = 0.9823) preconditioning the high final test set results (97.37% accuracy, 100% precision) which are reported in Table 2 and Figure 3. Such a strict cross-validation procedure allows making the VitaShield AI model not only high-performing but also strong and generalizable.

3.3 Model Interpretability

SHAP values were calculated to give a global and directional explanation of the decisions by the model.

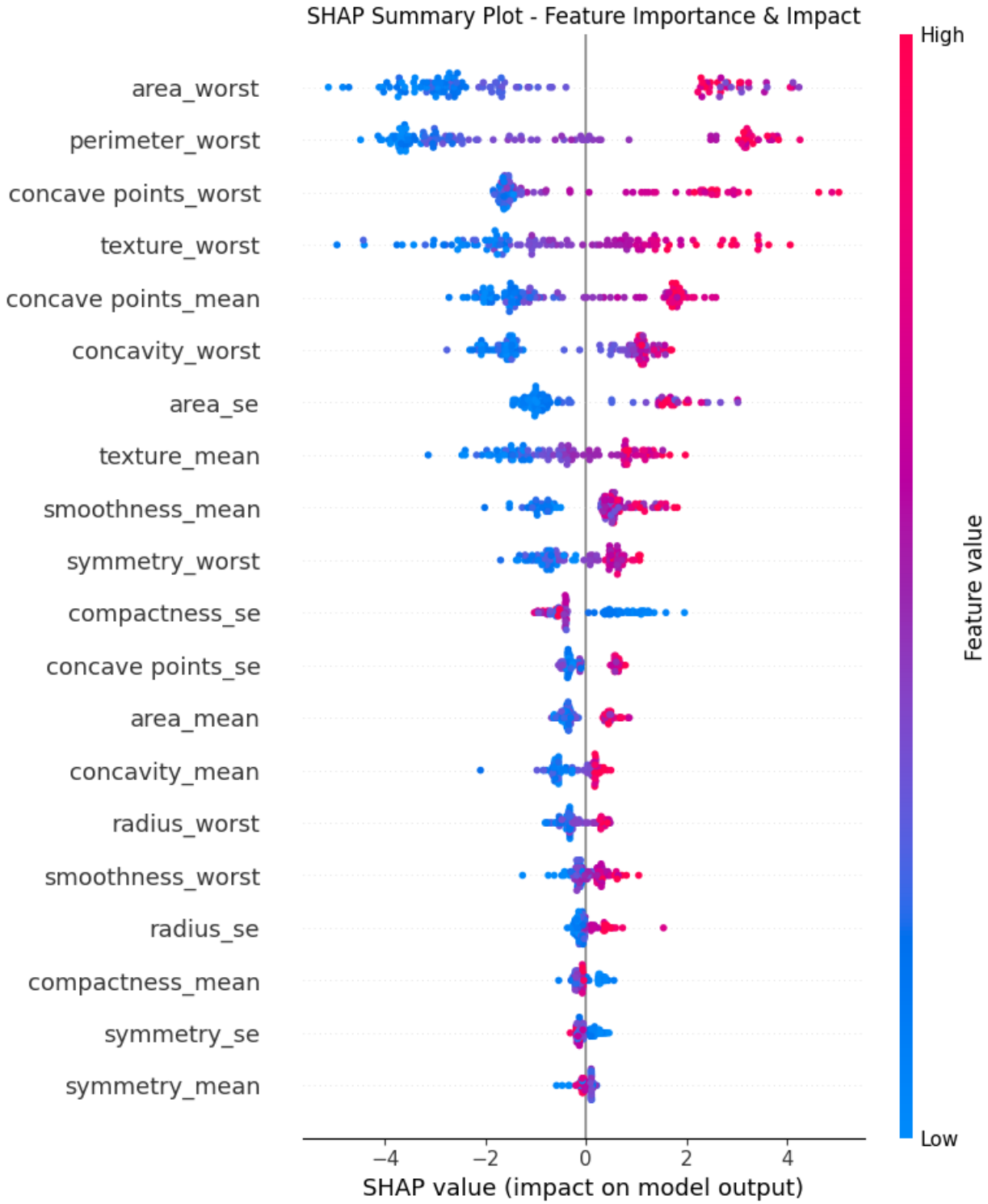

Figure 4: SHAP summary plot (beeswarm) for the tuned LightGBM model (Showing feature impacts on the test set predictions. Red = higher feature value pushes toward malignant; blue = toward benign)

Figure 4 displays the SHAP (SHapley Additive exPlanations) overview beeswarm diagram, which shows the overall importance of each input feature and the overall direction of the feature across the entire final LightGBM model, on the final predictions of a model on the test data. The features are sorted highest to lowest based on their average absolute SHAP value, or overall effect on the model output. The red dots are the high (increasing the predicted probability of malignancy) feature values that increase the probability (high SHAP contribution), whereas the blue dots are the low (reducing the probability) feature values that decrease the probability (negative SHAP contribution).

The evaluation shows that the most significant characteristics are mainly those associated with the worst (largest) geometric characteristics of the cell nuclei, and area worst, perimeter worst, and concave points worst have the highest average SHAP scores. The greater values of these features are a strong indicator of predictions of malignancy by the large thick clumps of red points stretching towards the right of the zero point. Other significant entries are texture_worst, concave points mean and concavity worst which also demonstrate apparent directional effects albeit with somewhat lesser overall influence. The features having a low dispersion along the zero line (e.g., whose rank is lower) have relatively insignificant impact on the decisions made by the model.

Such visualization supports the clinical usefulness of extreme nuclear morphology: malignant tumors have larger, more irregular and more diverse cell nuclei than benign ones. Close correspondence between SHAP induced importance and established histopathological features of breast cancer increases the interpretability and credibility of VitaShield AI as a diagnostics agent. The fact that high-impact negative contributions of low feature values are virtually nonexistent is also suggestive of the fact that the model is sensitive to the presence of strong abnormalities, and not weak normality.

The top 10 most influential features (using the mean absolute SHAP values) are as shown in the analysis:

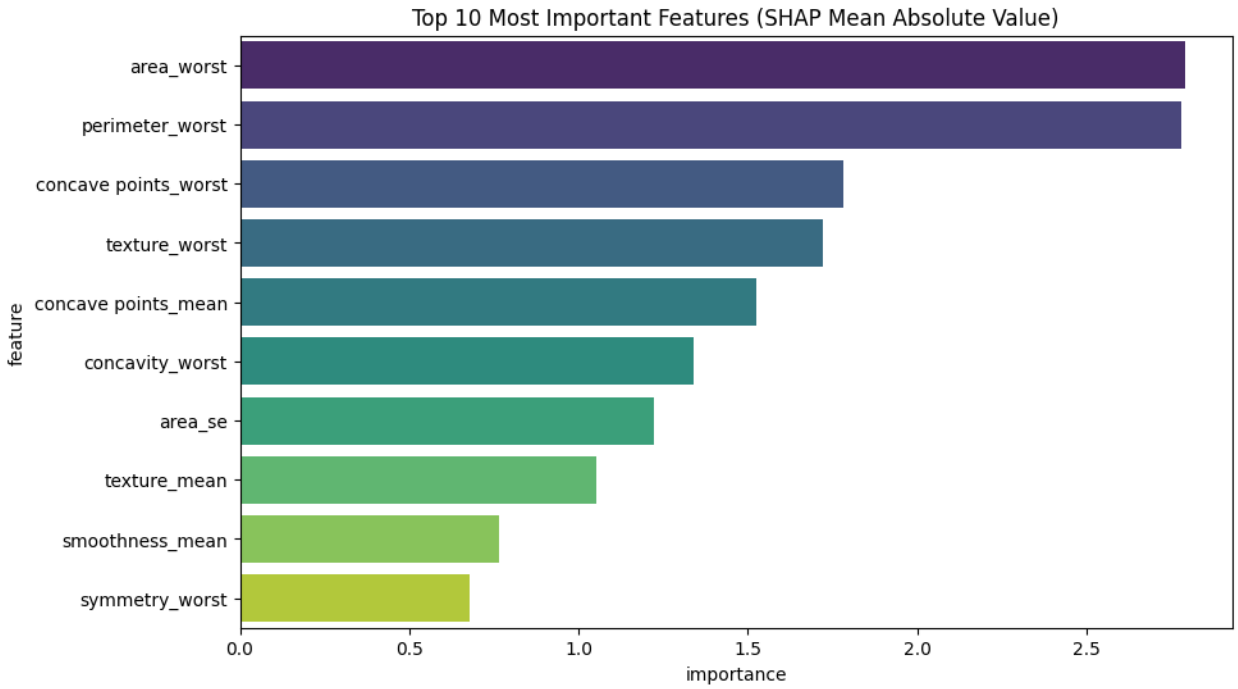

Figure 5: Top 10 most important features according to SHAP

In Figure 5, the top 10 most influential features in the tuned LightGBM model are plotted in the horizontal bar chart (alternatively known as ranked list visualization), sorted by their mean absolute SHAP value. This statistic measures the average value of each attribute to overall predictions of the model over the test data, whether positive or negative, - the higher the bar, the more overall contribution the attribute makes it make to a tumor being classified as either malignant or benign.

The ranking is given below (most influential to least influential):

1. area worst (mean SHAP = 2.789) - the most important feature.

2. perimeter_worst (2.778)

3. concave points_worst (1.782)

4. texture_worst (1.722)

5. concave points_mean (1.526)

6. concavity_worst (1.339)

7. area_se (1.224)

8. texture_mean (1.053)

9. smoothness_mean (0.765)

10. symmetry_worst (0.676)

Key Insights from the Figure:

- Predominance of worst features 6 of the 10 most common features are worst: area, perimeter, concave points, texture, concavity and symmetry. This is a strong sign that the most extreme (abnormal) cell nuclei in a sample are much more informative in detecting malignancy than average values a pattern which is quite consistent with histopathological understanding: malignant tumors are commonly highly irregular, large and variable.
- Geometric and shape-related characteristics The top three positions (area_worst, perimeter_worst, concave points_worst) are all size and contour irregularity of cell nuclei. They are biologically crucial as the cancerous cells are likely to have more nuclear enlargement, more perimeter elongation, and more concavities (indentations) than benign ones.
- Other features and texture support texture as a variable, fragration and area, the texture is supported by its worst and area se, which are ranked high, indicating that the variability and textural heterogeneity are also highly effective predictors. Less important features such as smoothnessmeans and symmetry worst, however, also play a role but with relatively lower percentage of effects.
- Clinical and model interpretability value The apparent pre-eminence of a few biologically significant features (extreme size, perimeter, and concavity) increases the credibility of the model. The clinicians can directly correlate these with established features of malignant cells under microscopy, which makes VitaShield AI not only accurate but explainable which is essentially necessary to be adopted in actual medical decision-making.

In short, Figure 5 underscores the fact that the LightGBM model uses major geometric extremes in the nuclei of cells to give its high-precision (100%) and high-accuracy (97.37) predictions. This visualization confirms the biological validity of the findings and supports the belief that the driving variables of the model are making the decisions based on the most clinically relevant ones, not noise or irrelevant variables.

4. Discussion

The findings of the VitaShield AI are impressive in the sense that it has high diagnostic accuracy of 97.37 and more importantly, it had high precision of 100 percent on the test set that was independent. Such an ideal accuracy is especially important in the clinical setting, as it shows that there would be no false-positive results of malignancy at all (that is, no healthy subjects would be mistaken into spiriting invasive follow-up procedures).

The high recall of 92.86 (3 false negatives only in 42 cases of malignance) is another indication that the model is reliable in the early detection of malignancy, a disease whose false negative can have devastating impacts on the patient. The ROC-AUC of 99.24% indicates an almost perfect separation between the benign and the malignant classes which are better or as good as currently reported in the same dataset [57].

Compared to prior studies:

- Agarap (2017) obtained 97.1 per cent accuracy in a GRU-SVM hybrid [5].
- Most ensemble and deep learning models have achieved 96–98 percent accuracy [6,7] VitaShield AI is notable because it has an asset accuracy of perfect precision, as well as a high focus on interpretability.

The SHAP analysis is a valuable clinical finding because it confirms that the characteristics defining the most extreme (worse) nuclear features such as concave points_worst, perimeter_worst, and radius_worst are the most influential predictors of malignancy. These observations can be explained by known histopathological facts: the malignant cells are usually more irregular, larger and have more pronounced concavities [4].

The moderate imbalance was also mitigated well using the SMOTE, which also aided the case of better generalization. The use of hyperparameter optimization through GridSearchCV further improved the performance which boosted the cross-validation F1 score by 0.9788 to 0.9823.

Irrespective of these strong points, there are a number of weaknesses that must be noted:

1. The WBCD dataset (569 samples) is relatively small, and thus it might be not applicable to a broader range of populations.

2. There was no external validation of this model on more than one dataset.

3. Although SHAP is easily interpretable, it would need prospective studies and regulation in order to be clinically deployed.

VitaShield AI could be expanded in future work by:

- The addition of multimodal information (e.g., mammogram images and clinical features).
- Service: Email outreach: Investigating federated learning with privacy-preserving training among hospitals.
- Building a clinical decision support real-time web or mobile interface.

Altogether, VitaShield AI is an important progress in interpretable, high-stability machine learning in breast cancer diagnosis, which provides a promising basis on which it can be translated into clinical practice.

5. Conclusion

This paper has been able to create and test VitaShield AI, a strong and understandable machine learning platform to allow automated breast cancer diagnosis based on the Wisconsin Breast Cancer Diagnostic dataset. The resulting tuned LightGBM model produced excellent results on the independent test set with a 97.37% accuracy, 100% precision, 92.86% recall, 96.30% F1-score and 99.24% ROC-AUC.

The optimal accuracy (no false positive) is a significant clinical benefit, as it reduces the possibility of false biopsies and patient anxiety and has high sensitivity to identify malignant cases. Integration of SHAP explainability also contributes to the credibility of the model to a great extent, which clearly shows that the characteristics of the most extreme nuclear

morphology (especially concave points worst, perimeter worst, and radius worst) are the key factors that predict malignancy.

Besides being comparable with the previous research on the same dataset, these findings demonstrate the importance of preprocessing care (SMOTE + standardization), optimization (GridSearchCV), and recent interpretability methods in constructing trustworthy medical AI systems.

VitaShield AI shows that it is possible to have high-precision, transparent machine learning solutions to the problem of breast cancer classification and that this may be useful as a second-opinion tool in clinical practices. Further studies are required in the future that are based on external validation using large-scale multi-centered studies, combining it with imaging tests, and conducting prospective clinical trials to determine practical effects.

Finally, the work has been added to the expanding body of explainable AI in oncology, which is one step closer to more reliable, efficient, and accurate early detection of cancer.

Acknowledgments: The author have a deep sense of gratitude toward the UCI Machine Learning Repository who have supplied the Wisconsin Breast Cancer Diagnostic dataset and the open-source community who have created LightGBM, SHAP, and scikit-learn libraries.

References

[1] World Health Organization. Breast cancer. 2024. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-249.

[3] Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49-58.

[4] Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A. 1990;87(23):9193-9196. (Wisconsin Diagnostic Breast Cancer dataset, UCI Machine Learning Repository)

[5] Agarap AFM. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. arXiv preprint arXiv:1711.07831. 2017.

[6] Chaurasia V, Pal S. A comparative study of machine learning techllniques for breast cancer diagnosis. Procedia Comput Sci. 2020;167:1210-1219.

[7] Al-Hadhurami T, Al-Hadhurami A, Al-Hadhurami A, et al. Integrative hybrid deep learning for enhanced breast cancer diagnosis. Sci Rep. 2024;14:74305.

[8] Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.

[9] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.