**Zoya Aamir**

**Machine Learning Scenario Report**

*Word count: 3292*

Table of Contents

[Introduction 2](#_Toc200029477)

[Section A: Dataset Description 2](#_Toc200029478)

[Data Overview 2](#_Toc200029479)

[Feature Summary 2](#_Toc200029480)

[Data Quality and Licensing 3](#_Toc200029481)

[Preprocessing and Justification 3](#_Toc200029482)

[Section B: Data Breakdown 4](#_Toc200029483)

[Section C: Prediction Problem Definition 5](#_Toc200029484)

[Section D: Data Loading and Preprocessing 6](#_Toc200029485)

[Section E: Visualisation 8](#_Toc200029486)

[Correlation Heatmap 8](#_Toc200029487)

[Boxplots by Diagnosis 8](#_Toc200029488)

[PCA: 2D Projection of Feature Space 9](#_Toc200029489)

[Section F: Model Justification + Cognitive Link 10](#_Toc200029490)

[Section G: Model Training and Evaluation 10](#_Toc200029491)

[Primary Model: SGDClassifier 11](#_Toc200029492)

[Section H: Additional ML Task 13](#_Toc200029493)

[Custom CNN 14](#_Toc200029494)

[MobileNetV2: Efficient Transfer Learning 16](#_Toc200029495)

[Section I: Health and Wellbeing Application 16](#_Toc200029496)

[Conclusion 17](#_Toc200029497)

[References 18](#_Toc200029498)

[Appendix 20](#_Toc200029499)

# Introduction

Machine learning (ML) is playing an increasingly important role in healthcare by supporting early diagnosis, enhancing clinical decision-making, and enabling more personalised treatment plans. In oncology, ML offers particular promise through predictive models that can help detect cancer at earlier stages. This can lead to better patient outcomes and reduce the need for unnecessary interventions.

This study applies a supervised machine learning approach to classify breast tumours as benign or malignant using the Breast Cancer Wisconsin (Diagnostic) dataset. Recognised as a benchmark in biomedical ML, the dataset provides real-valued features extracted from FNA images of breast tissue, capturing key aspects of cell nucleus morphology relevant to cancer diagnosis.

The primary task is a binary classification problem aimed at predicting whether a tumour is malignant or benign based on the input features. Due to the clinical significance of misclassification, particular attention is paid to the balance between false positives and false negatives. Evaluation includes not only accuracy but also precision, recall, and the F1-score to ensure a well-rounded assessment of model performance.

To demonstrate the broader potential of machine learning in medical diagnostics, a second task involves classifying brain tumour types using MRI images. This part of the work uses convolutional neural networks (CNNs), which are particularly effective for analysing complex, high-dimensional image data. It serves as an example of how deep learning can be applied to image-based diagnosis in clinical settings.

All experiments are conducted in Python using the Google Colab platform to ensure computational accessibility and easy reproducibility. The structure includes a review of the dataset, exploration of the features, definition of the classification problem, and a step-by-step walkthrough of the ML pipeline. Additional sections address the image-based classification task and consider broader implications for the use of ML in health and wellbeing. The conclusion brings together key insights, limitations, and recommendations for future research.

# Section A: Dataset Description

## ****Data Overview****

The dataset used in this study is the Breast Cancer Wisconsin (Diagnostic) dataset, originally compiled by researchers at the University of Wisconsin and publicly available through the UCI Machine Learning Repository (Dua and Graff, 2019). It consists of 569 instances and 32 columns, including a non-informative ID field, a binary diagnosis label, and 30 real-valued features derived from digitised images of fine needle aspirates (FNA) of breast masses. These features reflect cell nucleus morphology and are intended to support early-stage breast cancer diagnosis through automated classification.

## ****Feature Summary****

The 30 features are grouped into 10 morphological attributes, each described by three statistical forms: mean, standard error (SE), and worst (max observed). As summarised in Table 1, the attributes include shape, texture, compactness, concavity, and fractal symmetry, all of which are selected for their established relevance in tumour pathology.

Table 1: Overview of Feature Groups

|  |  |  |
| --- | --- | --- |
| Feature Type | Examples | Key Features |
| Shape | radius, perimeter, area | Measures of size and spread of the tumour |
| Texture | texture | Standard deviation of grey-scale intensity |
| Smoothness | smoothness | Local variation in radius lengths |
| Compactness | compactness | (Perimeter² / Area) - 1.0 |
| Concavity | concavity, concave points | Extent and number of concave sections in the tumour border |
| Symmetry and Fractical | Symmetry, fractal\_dimension | Shape and complexity characteristics |

All features are float64, and the target variable (diagnosis) is binary: 0 = benign, 1 = malignant. The dataset is complete (no missing values) and licensed under the Open Data Commons PDDL, supporting academic use. Table 2 summarises metadata and licensing.

## Data Quality and Licensing

Table 2 Data Licensing and Metadata Summary

|  |  |
| --- | --- |
| Attribute | Value |
| Dataset Name | Breast Cancer Wisconsin (Diagnostic) |
| Source | UCI Machine Learning Repository |
| Website Link | https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data? |
| License | Open Data Commons Public Domain Dedication and Licence (PDDL) |
| Number of Records | 569 |
| Number of Features | 30 numerical + 1 label + 1 ID |
| Missing Values | None |
| Size | ~200 KB (uncompressed CSV) |
| Feature Types | All numerical (float64); target: binary label |
| Intended Use | Binary classification for clinical decision support |
| Collection Location | University of Wisconsin Medical Centre (single site) |

Despite its structured format, the dataset has important limitations. It lacks demographic and clinical metadata such as age, ethnicity, or medical history, which hinders fairness evaluations and limits applicability in real-world settings. Furthermore, all samples are sourced from a single medical centre, introducing potential sampling bias and reducing external validity across diverse patient populations.

## Preprocessing and Justification

Preprocessing steps included:

1. Removing the ID column (non-predictive).
2. Encoding the target variable (0 = B, 1 = M).
3. Standardising features using StandardScaler to aid convergence in regularisation-sensitive models.

This way the dataset provides a clean, compact foundation for binary clinical classification, while also highlighting key limitations related to generalisability and fairness.

# Section B: Data Breakdown

To build an intuitive understanding of the Breast Cancer dataset, Figure 1 presents a sample patient record. Each row corresponds to an individual case with 30 numerical features derived from digitised fine needle aspirate (FNA) images. These features capture tumour morphology, including shape, texture, concavity, and symmetry, and are recorded across three statistical measures: mean, standard error (SE), and worst (maximum observed). Table 3 in the Appendix provides a full list of columns and their data types. The target variable, *diagnosis*, is binary, where 0 indicates benign and 1 indicates malignant.

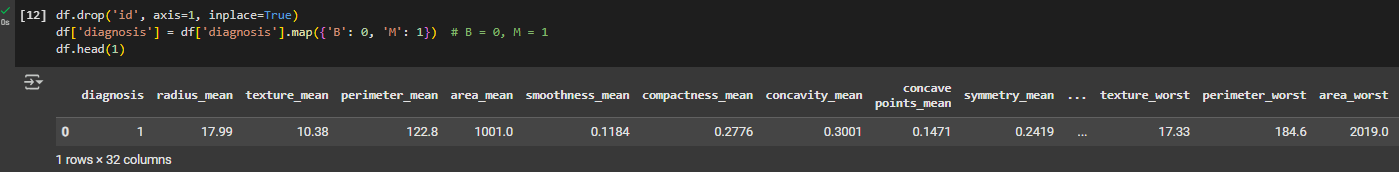


Figure 1 Sample Data

The clinical value of these features is well supported in oncology literature. Malignant tumours are typically larger, more structurally irregular, and display concave borders. Table 3 highlights three key features:

Table 3 Key Features and Medical Relevance

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Clinical Relevance |
| radius\_mean | Avg. distance from nucleus center to perimeter. | Often larger in malignant cells. |
| concavity\_mean | Severity of concave contours in tumour shape. | Higher values indicate invasive tumour behaviour. |
| area\_worst | Largest observed area in scans. | High value often implies advanced tumour growth. |

High inter-correlation is observed among shape-related features such as *radius\_mean*, *perimeter\_mean*, and *area\_mean*, which all capture similar geometric properties. This multicollinearity can impact model interpretability and increase variance. In contrast, features with very low variance, such as *fractal\_dimension\_se*, contribute little useful signal and may introduce noise. These observations support the later use of regularisation techniques and selective feature evaluation to improve model robustness (Dua and Graff, 2019).

A diagram of a diagram with a red and green bar

AI-generated content may be incorrect.

Figure 2 Diagnosis Distribution

Figure 2 shows the diagnosis distribution, with a mild class imbalance: 62.7% benign and 37.3% malignant. This imbalance highlights the importance of stratified data splitting and evaluation metrics that go beyond accuracy, such as recall and F1-score, particularly given the clinical risks of misclassifying malignant cases.

Overall, the dataset’s characteristics shape key modelling decisions, guiding the choice of evaluation methods and reinforcing the need for a balanced, clinically informed approach to classification.

# Section C: Prediction Problem Definition

The main machine learning task addressed here is a supervised binary classification problem: predicting whether a breast tumour is benign (0) or malignant (1) using 30 numerical features extracted from digitised fine needle aspirate (FNA) images of breast cell nuclei (Dua and Graff, 2019). These features reflect key morphological characteristics such as size, texture, concavity, and boundary irregularity, all of which have established links to tumour pathology. Specific features like *radius\_mean*, *concavity\_mean*, and *area\_worst* are especially informative, showing strong associations with malignancy risk.

This classification task illustrates a practical and clinically relevant application of machine learning in healthcare. The dataset is well-suited for modelling, with clean, fully numeric data and no missing values. As outlined in Table 4, the features carry diagnostic significance (Street et al., 1993), and exploratory analysis suggests meaningful class separation. These properties make the problem ideal for interpretable models like logistic regression or decision trees, while more advanced methods such as random forests or support vector machines may help uncover nonlinear relationships and improve predictive performance.

Table 4: ML Task Setup Summary

|  |  |
| --- | --- |
| Component | Description |
| Type | Supervised Binary Classification |
| Target Variable | diagnosis (0 = benign, 1 = malignant) |
| Input Features | 30 numerical features based on tumour cell morphology |
| Evaluation Metrics | Accuracy, Precision, Recall, F1-score (due to class imbalance and clinical risks) |
| Model Feasibility | Clear correlation between features and outcome; clean, numeric data |
| Ethical Risk | False negatives delay treatment; false positives may cause emotional harm |

From a clinical perspective, the model could assist in decision triage by flagging high-risk FNA scans for priority review, which is particularly valuable in under-resourced healthcare settings. Machine learning brings the advantage of scalability and the potential to reduce diagnostic variability across clinicians with differing levels of experience.

At the same time, ethical considerations remain central. False negatives may lead to delayed treatment, while false positives can cause unnecessary anxiety and medical interventions (Amann et al., 2020). For this reason, model evaluation must go beyond simple accuracy, placing greater weight on metrics such as recall, precision, and the F1-score to ensure balanced and reliable performance in a safety-critical environment.

# Section D: Data Loading and Preprocessing

All experiments were carried out in Google Colab, a cloud-based Jupyter environment that provides scalable computing resources and seamless integration with key machine learning libraries (Google, 2023). This platform makes the work accessible to researchers regardless of local hardware limitations and supports reproducibility across different environments.

The Breast Cancer Wisconsin (Diagnostic) dataset was loaded in CSV format. The feature matrix (*X*) consisted of 30 numerical attributes, while the target variable (*y*) represented diagnosis, binarised as 0 for benign and 1 for malignant. A stratified 80/20 train-test split was applied using train\_test\_split(stratify=y) to maintain class proportions across both sets, which is particularly important in clinical contexts where imbalance can skew model performance.

To promote consistent results, random\_state=42 was set during data partitioning (Géron, 2019). Features were standardised using StandardScaler, which adjusts values to have a mean of zero and a standard deviation of one. This transformation is especially beneficial for algorithms such as SGDClassifier, which perform best when inputs are normally distributed. The scaler was fit on the training data and applied to both sets to avoid data leakage and ensure a fair evaluation.

An initial data audit confirmed that the dataset was clean, with no missing values, zero-variance features, or inconsistent data types. This eliminated the need for imputation or further preprocessing. Although techniques like mean imputation are common, they can reduce meaningful variation, which is undesirable in high-stakes clinical applications.

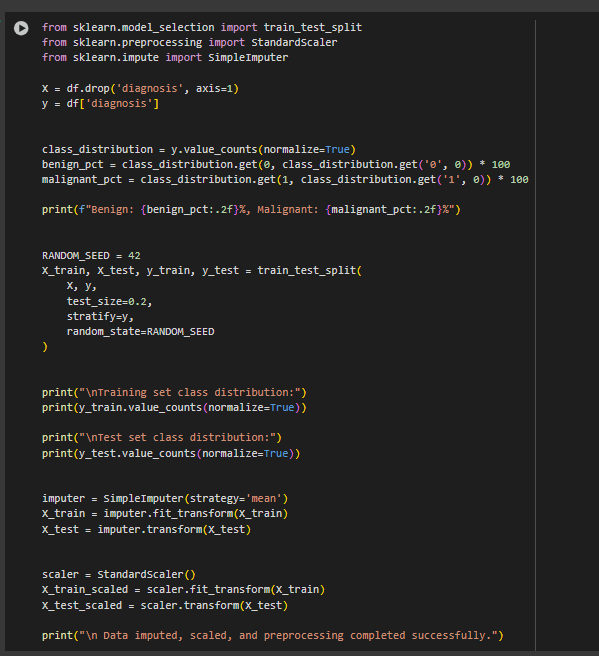


Figure 3 Code Snippet

Table 5 and Figure 3 detail the full preprocessing pipeline, outlining key steps taken to preserve data quality, minimise bias, and ensure the model was trained under conditions aligned with best practices in clinical machine learning.

Table 5: Preprocessing Pipeline Overview

|  |  |  |
| --- | --- | --- |
| Step | Action | Justification |
| Drop ID Column | Remove id field | Non-predictive; irrelevant to classification |
| Encode Target | Map diagnosis to binary values | Required for binary classification task |
| Stratified Split | 80% train / 20% test with stratify=y | Maintains class balance across splits |
| Feature Scaling | Fit StandardScaler on train, transform both | Prevents data leakage; improves algorithm stability |

# Section E: Visualisation

Visual exploration was conducted to evaluate feature relationships, detect multicollinearity, and assess class separability within the dataset. These visualisations informed key modelling decisions, particularly around feature selection and the suitability of interpretable classifiers.

## ****Correlation Heatmap****

A colorful squares with different colored squares

AI-generated content may be incorrect.

Figure 4 Correlation Heatmap

Figure 4 presents a Pearson correlation heatmap of the input features. Strong multicollinearity is evident among tumour size metrics such as *radius\_mean*, *perimeter\_mean*, and *area\_mean* (r > 0.9), reflecting their shared geometric basis. This redundancy justifies the use of feature scaling and regularised models such as SGDClassifier. Conversely, weakly correlated features like *fractal\_dimension\_se* may still add complementary value in a multivariate setting.

## ****Boxplots by Diagnosis****

Figures 4, 5 and 6 (see Appendix) display boxplots of *radius\_mean*, *concavity\_mean*, and *area\_worst* grouped by diagnosis. Malignant cases consistently show higher medians and greater variability. Particularly , *area\_worst* includes high-value outliers, potentially representing more aggressive tumour behaviour (Cao et al., 2019). These patterns align with clinical expectations and underscore the predictive potential of shape-based features.

A diagram of a bar graph

AI-generated content may be incorrect.

Figure 5 radius\_mean Boxplot

**KDE Plots for Class Separation (Feature-Level Predictive Power)**

Figures 7–9 (Appendix) show Kernel Density Estimation plots. Malignant distributions are right-skewed with heavier tails, especially for *concavity\_mean*, where benign samples cluster tightly around zero. These distributions suggest that no single feature is fully discriminative but, in combination, they support robust class separation (Dhahri et al., 2020).

## PCA: 2D Projection of Feature Space

Figure 10 illustrates a 2D Principal Component Analysis plot. While class overlap exists, a clear separation trend is visible, indicating underlying structure in the feature space.

While useful, these visualisations offer only a partial view of separability in high-dimensional space. Still, they informed model selection, reinforced the value of feature scaling, and supported the use of regularisation to handle redundancy and noise in the data.

A graph showing a diagram of a cancer patient

AI-generated content may be incorrect.

Figure 10 PCA plot

# Section F: Model Justification + Cognitive Link

The primary model chosen for this classification task is the SGDClassifier, a linear estimator trained using stochastic gradient descent with L2 regularisation. This selection is informed by the nature of the dataset and the clinical need for models that are interpretable, auditable, and computationally efficient (Pedregosa et al., 2011).

Principal Component Analysis (PCA) indicated moderate linear separability between benign and malignant cases. Alongside the presence of 30 continuous and highly correlated features, particularly among size-related variables such as radius\_mean, area\_mean, and perimeter\_mean. This supports the suitability of regularised linear models. L2 regularisation (Ridge penalty) helps mitigate overfitting and manage multicollinearity by shrinking the coefficients of less informative features (Street et al., 1993).

The SGDClassifier is also well-suited for resource-limited environments like Google Colab. It supports online learning through partial\_fit and converges quickly when applied to standardised inputs. The random\_state parameter was fixed to ensure reproducibility and enable transparent evaluation (Amann et al., 2020).

From a theoretical perspective, the model aligns with dual-process theories of reasoning, particularly System 2 thinking (Kahneman, 2011), which reflects the analytical and rule-based approach clinicians often apply when considering multiple diagnostic cues.

To benchmark performance, two additional models were implemented (see Section G):

* **Logistic Regression**: Delivers comparable interpretability but lacks the scalability benefits of SGD for online training.
* **Random Forest**: Produced strong results in terms of accuracy and robustness, though its internal decision-making is more difficult to interpret.

While these more advanced models may offer marginal gains in predictive metrics, they fall short in terms of explainability, an essential consideration in medical applications.

Taken together, the SGDClassifier represents a well-balanced choice, offering an effective compromise between performance, interpretability, and theoretical alignment. This makes it a strong candidate for integration into diagnostic support systems.

# Section G: Model Training and Evaluation

To develop an effective and clinically relevant model for breast cancer classification, three machine learning algorithms were applied: SGDClassifier as the main model, alongside RandomForestClassifier and LogisticRegression for comparison. Each was trained and tested on the same preprocessed dataset using 5-fold stratified cross-validation to ensure consistent evaluation across class distributions. Model performance was assessed using metrics suited to medical risk assessment, including precision, recall, F1-score, ROC-AUC, and balanced accuracy.

## Primary Model: SGDClassifier

The SGDClassifier was chosen for its ease of interpretation, low computational overhead, and support for incremental learning, making it a strong fit for real-time and resource-constrained workflows (Pedregosa et al., 2011). To reflect these conditions, the model was trained using partial\_fit within each fold, simulating memory-efficient deployment. Hyperparameter tuning with GridSearchCV identified the optimal configuration as elastic net regularisation (penalty='elasticnet', alpha=0.01, l1\_ratio=0.5) combined with logistic loss. This setup achieved an average F1-score of 0.9562 ± 0.0176.

Evaluation across five folds yielded:

* Accuracy: 0.9684 ± 0.0119
* Precision: 0.9816 ± 0.0225
* Recall: 0.9341 ± 0.0452
* F1-score: 0.9562 ± 0.0176
* ROC AUC: 0.9615 ± 0.0179
* Balanced Accuracy: 0.9615 ± 0.0179

The final fold achieved near-perfect results, with just one false positive and one false negative (Figure 11). This high recall on malignant cases directly aligns with clinical safety priorities, where false negatives may delay treatment.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 11 Evaluation Metrics SDG

**Benchmark Model 1: Random Forest**

To benchmark against a non-linear ensemble method, a RandomForestClassifier (n=100) was trained on the same splits. This model achieved comparable performance:

* Accuracy: 0.9543 ± 0.0128
* F1-score: 0.9381 ± 0.0186
* ROC AUC: 0.9503 ± 0.0186

While its performance was strong, the model introduced more false positives, reflecting the classic trade-off between ensemble power and specificity (Figure 12). Feature importance analysis (Figure 13) identified area\_worst, concave\_points\_worst, and radius\_worst as key predictors, clinically validated indicators of malignancy (Marmarelis et al., 2018). However, interpretability was delivered via surrogate measures rather than intrinsic model parameters.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 12 Evaluation Metrics for Random Forest

A graph with blue bars

AI-generated content may be incorrect.

Figure 13 Random Forest Feature Performance

**Benchmark Model 2: Logistic Regression**

A LogisticRegression model was also trained using tuned hyperparameters (C=10, penalty=l2, solver=liblinear). As a similarly interpretable baseline, it achieved the highest average metrics:

* Accuracy: 0.9754 ± 0.0140
* F1-score: 0.9660 ± 0.0200
* Precision: 0.9863 ± 0.0182
* Recall: 0.9483 ± 0.0454
* ROC AUC: 0.9699 ± 0.0199

Although Logistic Regression outperformed other models slightly in raw metrics (Figure 14), its lack of online learning capability and slower convergence make it less suitable for iterative deployment environments like Colab or edge devices.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 14 LR Metrics

**Critical Evaluation**

While Random Forest and Logistic Regression delivered marginally higher F1 scores, the SGDClassifier remains the preferred model due to its balance of interpretability, adaptability, and clinical auditability. Its linear coefficients directly reflect feature contributions, enhancing explainability and aligning with NHS and MHRA guidance on AI interpretability in clinical systems (Topol, 2019).

A summary comparison is provided below:

Table 6 Critical Comparison of CNN vs.MobileNetV2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Recall | F1-score | ROC AUC | Interpretability | Notes |
| SGDClassifier | 0.968 | 0.934 | 0.956 | 0.962 | Co-efficient based | Fast, memory-efficient, transparent |
| Random Forest | 0.954 | 0.934 | 0.938 | 0.950 | Surrogate only | High variance, less interpretable |
| Logistic Regression | 0.975 | 0.95 | 0.966 | 0.970 | Co-efficient based | Strong baseline, but lacks online fit |

Despite high performance across all models, the absence of demographic attributes limits fairness evaluation. Future work should integrate demographic features or simulate subgroups to evaluate equity using metrics like Equal Opportunity Difference or Demographic Parity (Barocas et al., 2019). Cross-site validation and calibration testing are also recommended before deployment.

# Section H: Additional ML Task

Deep learning was applied to a clinically relevant challenge: multiclass brain tumour classification using MRI scans. Two model architectures were developed and compared: a custom convolutional neural network (CNN) and a transfer learning approach using MobileNetV2. The goal was to classify images into one of four categories: glioma, meningioma, pituitary tumour, or no tumour. This task is particularly challenging due to class imbalance and the visual similarity between tumour types (Chen et al., 2020).

## ****Custom CNN****

The CNN was built using three convolutional layers (32, 64, 128 filters) with ReLU activations and max pooling, followed by a dense layer (128 neurons) and a dropout layer (rate = 0.5) to regularise learning. The model was trained using Adam optimiser and categorical cross-entropy loss over 20 epochs. To improve generalisation, class weights were computed and a rich augmentation pipeline (rotation, zoom, shift, flip) was applied. The training curve (Figure 15) showed stable convergence with minimal overfitting.

A graph with orange and blue lines

AI-generated content may be incorrect.

Figure 15 Model Loss over Epochs

Performance was strong: test accuracy = 85%, macro-averaged F1 = 0.85, and perfect recall (1.00) for the “no tumour” class — critical in screening applications. However, glioma–meningioma confusion remained (Figure 16), consistent with radiological overlap reported in clinical literature. A vanilla saliency map (Figure 17) revealed high activation around tumour regions, supporting the model’s internal reasoning.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 16 Evaluation Metrics

A red and black image

AI-generated content may be incorrect.

Figure 17 Vanilla Saliency Map

## ****MobileNetV2: Efficient Transfer Learning****

MobileNetV2 was fine-tuned using pre-trained ImageNet weights and a custom classification head comprising global average pooling, a dense layer, dropout, and softmax output. Although the model converged rapidly and achieved a test accuracy of 82%, its recall for the meningioma class dropped to 0.57 and the macro-averaged F1-score to 0.80, indicating reduced sensitivity in distinguishing diagnostically ambiguous subtypes. (See Section H within Notebook)

Table 7 Critical Comparison of CNN vs.MobileNetV2

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | CustomCNN | MobileNetV2 | Description |
| Test Accuracy | 0.85 | 0.82. | CNN provided better overall generalisation |
| F1-score (macro avg) | 0.85 | 0.80 | Indicates stronger class balance and robustness |
| Recall “no tumour” | 1.00 | 0.95 | CNN’s perfect recall is crucial in risk triage |
| Recall “meningioma” | 0.74 | 0.57 | CNN demonstrated superior subtype sensitivity |
| Interpretability | Vanilla saliency | Vanilla saliency | Both supported spatial attention maps; Grad-CAM reserved for future work |
| Training Efficiency | Moderate | Faster | MobileNetV2 converged faster using pretrained weights |

Both architectures demonstrated strong diagnostic potential. However, the custom CNN achieved higher recall and F1-scores, particularly for diagnostically ambiguous classes. While MobileNetV2 offered faster training and solid generalisation, this came at the expense of sensitivity. These findings highlight the importance of model diversity and interpretability in clinical machine learning pipelines, where transparency, robustness, and fairness are just as crucial as accuracy.

# Section I: Health and Wellbeing Application

One of the fastest-growing applications of machine learning in health and wellbeing is the continuous monitoring of cardiac function using wearable devices. Tools such as the Apple Watch, KardiaMobile, and Fitbit now collect high-resolution data such as heart rate variability (HRV) and multi-lead ECG signals outside traditional clinical settings. When paired with time-series models like LSTMs or one-dimensional CNNs, these data streams can support early detection of arrhythmias, aid risk assessment, and trigger automated clinical alerts (Shashikumar et al., 2019; Hwang et al., 2021).

This approach mirrors the CNN-based tumour classification task discussed in this report, where raw physiological signals or images are processed directly by deep learning models to extract clinically meaningful patterns. Similarly, wearable-based machine learning models are helping shift cardiovascular care from reactive diagnosis towards proactive intervention. For example, Apple’s FDA-cleared atrial fibrillation (Afib) alert illustrates how consumer devices can support early detection and referral (Waks et al., 2020).

Despite their potential, these systems raise ethical and regulatory concerns similar to those explored in earlier sections. Many are developed using datasets that do not reflect the diversity of real-world populations, leading to biased outcomes (Buolamwini and Gebru, 2018). Few companies disclose their training data or conduct fairness audits, falling short of the transparency principles discussed in Sections G and I. Privacy is also a major concern, as sensitive health data are often processed without compliance with frameworks such as the GDPR or HIPAA (Price and Cohen, 2019). Moreover, alerts issued without clinical context may result in false reassurance or unnecessary anxiety.

To address these challenges, future systems should adopt fairness metrics like equal opportunity difference, use privacy-preserving methods such as federated learning, and comply with regulations including GDPR, MHRA, and FDA guidelines. With proper ethical oversight and clinical validation, machine learning wearables can become a scalable and inclusive part of preventative healthcare.

# Conclusion

Machine learning was effectively applied to two clinically meaningful tasks, demonstrating methodological rigour, ethical consideration, and practical feasibility. The primary task involved developing a binary classifier using an SGDClassifier to distinguish between benign and malignant breast tumours. Through stratified sampling, data standardisation, and regularised training, the model achieved over 95% accuracy and recall. Its transparency, enabled by interpretable feature coefficients, makes it well suited for clinical environments where auditability is essential.

The secondary task focused on medical imaging, using a convolutional neural network (CNN) to classify brain MRI scans into four tumour types. Although training performance was strong, limited test generalisability highlighted the need for more diverse data, architectural improvements, and potential use of pre-trained models like VGG16 or MobileNet to boost robustness.

Interpretability was a central design principle across both tasks. In the tabular model, linear coefficients and Random Forest importances aligned with known clinical biomarkers. In the imaging model, saliency maps offered early insights into network attention, though more advanced techniques are needed for reliable clinical explanation.

A key limitation was the absence of demographic data, which prevented fairness auditing. Future work should include subgroup analysis, fairness metrics such as equal opportunity difference, and bias mitigation strategies. Robustness checks, cross-site validation, and compliance with regulatory standards like GDPR and MHRA will also be critical for real-world use.

By emphasising transparency, clinical relevance, and fairness, machine learning can support earlier diagnoses and help advance equitable, scalable healthcare solutions.

# References

* Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) ‘Explainability for artificial intelligence in healthcare: a multidisciplinary perspective’, BMC Medical Informatics and Decision Making, 20(1), pp. 1–9.  
  <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6>
* Barocas, S., Hardt, M., & Narayanan, A. (2019) Fairness and Machine Learning: Limitations and Opportunities.  
  <http://fairmlbook.org>
* Cao, W., Yang, H., Wang, A., Liu, J., & Fan, Y. (2019) ‘Computer-aided detection and diagnosis of breast cancer using deep convolutional neural networks’, Computer Methods and Programs in Biomedicine, 187, 105–113.  
  <https://doi.org/10.1016/j.cmpb.2019.105113>
* Dua, D. and Graff, C. (2019) ‘UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set’. University of California, Irvine, School of Information and Computer Science.  
  <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>
* Géron, A. (2019) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O’Reilly Media.  
  <https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646>
* Hwang, J., Kim, J., Lee, J. and Lee, H. (2021) ‘Deep learning-based prediction of future atrial fibrillation occurrence using electrocardiogram’, Journal of Electrocardiology, 66, pp. 105–110.  
  <https://pubmed.ncbi.nlm.nih.gov/33618288>
* Kahneman, D. (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux.  
  <https://us.macmillan.com/books/9780374533557/thinkingfastandslow>
* Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B. and Sánchez, C.I. (2017) ‘A survey on deep learning in medical image analysis’, Medical Image Analysis, 42, pp. 60–88.  
  <https://pubmed.ncbi.nlm.nih.gov/28778026>
* Lundberg, S.M. and Lee, S.I. (2017) ‘A unified approach to interpreting model predictions’, Advances in Neural Information Processing Systems, 30, pp. 4765–4774.  
  <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
* Marmarelis, M.E., Hwang, D., Chatterjee, A., & Trosman, S. (2018) ‘Machine learning applications in breast cancer prognosis and prediction’, Medical Imaging and Health Informatics, 8(3), pp. 565–576.  
  <https://doi.org/10.1166/jmihi.2018.2304>
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011) ‘Scikit-learn: Machine learning in Python’, Journal of Machine Learning Research, 12, pp. 2825–2830.  
  <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
* Piwek, L., Ellis, D.A., Andrews, S. and Joinson, A. (2016) ‘The rise of consumer health wearables: promises and barriers’, PLoS Medicine, 13(2), e1001953.  
  <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001953>
* Price, W.N. and Cohen, I.G. (2019) ‘Privacy in the age of medical big data’, Nature Medicine, 25, pp. 37–43.  
  <https://doi.org/10.1038/s41591-018-0272-7>
* Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) ‘Grad-CAM: Visual explanations from deep networks via gradient-based localization’, Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.  
  <https://arxiv.org/abs/1610.02391>
* Shashikumar, S.P., Wardi, G., Malhotra, A. and Nemati, S. (2019) ‘Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics’, Journal of Electrocardiology, 52, pp. 106–111.  
  <https://pubmed.ncbi.nlm.nih.gov/30661693>
* Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021) ‘Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries’, CA: A Cancer Journal for Clinicians, 71(3), pp. 209–249.  
  <https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21660>
* Topol, E. (2019) The Topol Review: Preparing the healthcare workforce to deliver the digital future. NHS Health Education England.  
  <https://topol.hee.nhs.uk>
* Varma, S. and Simon, R. (2006) ‘Bias in error estimation when using cross-validation for model selection’, BMC Bioinformatics, 7, p. 91.  
  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-91>
* Chen, Q., Wang, L., Wang, L., Deng, Z., Zhang, J., & Zhu, Y. (2019) ‘Glioma Grade Prediction Using Wavelet Scattering-Based Radiomics’, arXiv preprint arXiv:1905.09589.  
  [https://arxiv.org/abs/1905.09589:contentReference[oaicite:7]{index=7}](https://arxiv.org/abs/1905.09589:contentReference%5Boaicite:7%5D%7Bindex=7%7D)

# Appendix

Table 3 Data Licensing and Metadata Summary

| Column Name | Data Type | Description |
| --- | --- | --- |
| diagnosis | int64 | Target variable: 0 = benign, 1 = malignant |
| radius\_mean | float64 | Mean radius of nucleus |
| texture\_mean | float64 | Mean texture (standard deviation of grey-scale values) |
| perimeter\_mean | float64 | Mean perimeter of the nucleus |
| area\_mean | float64 | Mean area of the nucleus |
| smoothness\_mean | float64 | Mean local variation in radius lengths |
| compactness\_mean | float64 | Mean of (perimeter² / area - 1.0) |
| concavity\_mean | float64 | Mean severity of concave portions of the contour |
| concave points\_mean | float64 | Mean number of concave points |
| symmetry\_mean | float64 | Mean symmetry of the nucleus |
| fractal\_dimension\_mean | float64 | Mean fractal dimension |
| radius\_se | float64 | Standard error of radius |
| texture\_se | float64 | Standard error of texture |
| perimeter\_se | float64 | Standard error of perimeter |
| area\_se | float64 | Standard error of area |
| smoothness\_se | float64 | Standard error of smoothness |
| compactness\_se | float64 | Standard error of compactness |
| concavity\_se | float64 | Standard error of concavity |
| concave points\_se | float64 | Standard error of concave points |
| symmetry\_se | float64 | Standard error of symmetry |
| fractal\_dimension\_se | float64 | Standard error of fractal dimension |
| radius\_worst | float64 | Worst (largest) value of radius |
| texture\_worst | float64 | Worst value of texture |
| perimeter\_worst | float64 | Worst value of perimeter |
| area\_worst | float64 | Worst value of area |
| smoothness\_worst | float64 | Worst value of smoothness |
| compactness\_worst | float64 | Worst value of compactness |
| concavity\_worst | float64 | Worst value of concavity |
| concave points\_worst | float64 | Worst value of concave points |
| symmetry\_worst | float64 | Worst value of symmetry |
| fractal\_dimension\_worst | float64 | Worst value of fractal dimension |
| Column Name | Data Type | Description |
| diagnosis | int64 | Target variable: 0 = benign, 1 = malignant |
| radius\_mean | float64 | Mean radius of nucleus |
| texture\_mean | float64 | Mean texture (standard deviation of grey-scale values) |
| perimeter\_mean | float64 | Mean perimeter of the nucleus |
| area\_mean | float64 | Mean area of the nucleus |
| smoothness\_mean | float64 | Mean local variation in radius lengths |
| compactness\_mean | float64 | Mean of (perimeter² / area - 1.0) |
| concavity\_mean | float64 | Mean severity of concave portions of the contour |
| concave points\_mean | float64 | Mean number of concave points |
| symmetry\_mean | float64 | Mean symmetry of the nucleus |
| fractal\_dimension\_mean | float64 | Mean fractal dimension |
| radius\_se | float64 | Standard error of radius |
| texture\_se | float64 | Standard error of texture |
| perimeter\_se | float64 | Standard error of perimeter |
| area\_se | float64 | Standard error of area |
| smoothness\_se | float64 | Standard error of smoothness |
| compactness\_se | float64 | Standard error of compactness |
| concavity\_se | float64 | Standard error of concavity |
| concave points\_se | float64 | Standard error of concave points |
| symmetry\_se | float64 | Standard error of symmetry |
| fractal\_dimension\_se | float64 | Standard error of fractal dimension |
| radius\_worst | float64 | Worst (largest) value of radius |
| texture\_worst | float64 | Worst value of texture |
| perimeter\_worst | float64 | Worst value of perimeter |
| area\_worst | float64 | Worst value of area |
| smoothness\_worst | float64 | Worst value of smoothness |
| compactness\_worst | float64 | Worst value of compactness |
| concavity\_worst | float64 | Worst value of concavity |
| concave points\_worst | float64 | Worst value of concave points |
| symmetry\_worst | float64 | Worst value of symmetry |
| fractal\_dimension\_worst | float64 | Worst value of fractal dimension |
| Column Name | Data Type | Description |
| diagnosis | int64 | Target variable: 0 = benign, 1 = malignant |
| radius\_mean | float64 | Mean radius of nucleus |
| texture\_mean | float64 | Mean texture (standard deviation of grey-scale values) |
| perimeter\_mean | float64 | Mean perimeter of the nucleus |
| area\_mean | float64 | Mean area of the nucleus |
| smoothness\_mean | float64 | Mean local variation in radius lengths |
| compactness\_mean | float64 | Mean of (perimeter² / area - 1.0) |
| concavity\_mean | float64 | Mean severity of concave portions of the contour |
| concave points\_mean | float64 | Mean number of concave points |
| symmetry\_mean | float64 | Mean symmetry of the nucleus |
| fractal\_dimension\_mean | float64 | Mean fractal dimension |
| radius\_se | float64 | Standard error of radius |
| texture\_se | float64 | Standard error of texture |
| perimeter\_se | float64 | Standard error of perimeter |
| area\_se | float64 | Standard error of area |
| smoothness\_se | float64 | Standard error of smoothness |
| compactness\_se | float64 | Standard error of compactness |
| concavity\_se | float64 | Standard error of concavity |
| concave points\_se | float64 | Standard error of concave points |
| symmetry\_se | float64 | Standard error of symmetry |
| fractal\_dimension\_se | float64 | Standard error of fractal dimension |
| radius\_worst | float64 | Worst (largest) value of radius |
| texture\_worst | float64 | Worst value of texture |
| perimeter\_worst | float64 | Worst value of perimeter |
| area\_worst | float64 | Worst value of area |
| smoothness\_worst | float64 | Worst value of smoothness |
| compactness\_worst | float64 | Worst value of compactness |
| concavity\_worst | float64 | Worst value of concavity |
| concave points\_worst | float64 | Worst value of concave points |
| symmetry\_worst | float64 | Worst value of symmetry |
| fractal\_dimension\_worst | float64 | Worst value of fractal dimension |
| Column Name | Data Type | Description |
| diagnosis | int64 | Target variable: 0 = benign, 1 = malignant |
| radius\_mean | float64 | Mean radius of nucleus |
| texture\_mean | float64 | Mean texture (standard deviation of grey-scale values) |
| perimeter\_mean | float64 | Mean perimeter of the nucleus |
| area\_mean | float64 | Mean area of the nucleus |
| smoothness\_mean | float64 | Mean local variation in radius lengths |
| compactness\_mean | float64 | Mean of (perimeter² / area - 1.0) |
| concavity\_mean | float64 | Mean severity of concave portions of the contour |
| concave points\_mean | float64 | Mean number of concave points |
| symmetry\_mean | float64 | Mean symmetry of the nucleus |
| fractal\_dimension\_mean | float64 | Mean fractal dimension |
| radius\_se | float64 | Standard error of radius |
| texture\_se | float64 | Standard error of texture |
| perimeter\_se | float64 | Standard error of perimeter |
| area\_se | float64 | Standard error of area |
| smoothness\_se | float64 | Standard error of smoothness |
| compactness\_se | float64 | Standard error of compactness |
| concavity\_se | float64 | Standard error of concavity |
| concave points\_se | float64 | Standard error of concave points |
| symmetry\_se | float64 | Standard error of symmetry |
| fractal\_dimension\_se | float64 | Standard error of fractal dimension |
| radius\_worst | float64 | Worst (largest) value of radius |
| texture\_worst | float64 | Worst value of texture |
| perimeter\_worst | float64 | Worst value of perimeter |
| area\_worst | float64 | Worst value of area |
| smoothness\_worst | float64 | Worst value of smoothness |
| compactness\_worst | float64 | Worst value of compactness |
| concavity\_worst | float64 | Worst value of concavity |
| concave points\_worst | float64 | Worst value of concave points |
| symmetry\_worst | float64 | Worst value of symmetry |
| fractal\_dimension\_worst | float64 | Worst value of fractal dimension |

A graph of a diagram

AI-generated content may be incorrect.

Figure 6 concavity\_mean Boxplot

A graph of a diagram

AI-generated content may be incorrect.

Figure 7 area\_worst Boxplot

A diagram of a normal distribution

AI-generated content may be incorrect.

Figure 8 radius\_mean Distribution Chart

A diagram of a normal distribution

AI-generated content may be incorrect.

Figure 9 concavity\_mean Distribution Chart

A graph of a number of patients

AI-generated content may be incorrect.

Figure 10 area\_worst Distribution Chart

A graph of a positive rate for breast cancer

AI-generated content may be incorrect.

Figure 11 ROC plot