Using Machine Learning and LLM for Classifying of Benign and Malignant Cells from Breast Cancer Dataset

Aqeel Ahmed Khan[1], Bushra Shaheen[2], Masroor Ahmed[1]
[1]Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, A.Q. Khan Institute of Computer Sciences & Information Technology (KICSIT), Kahuta, Pakistan
*Correspondence: aakj666@gmail.com

Citation | Khan. A. A, Shaheen. B, Ahmed. M, "Using machine learning and LLM for classifying of benign and malignant cells from breast cancer dataset", IJIST, Vol. 7 Issue. 11 pp 190-202, December 2025
Received | October 29, 2025 **Revised |** November 29, 2025 **Accepted |** December 03, 2025 **Published |** December 07, 2025.

The most frequently diagnosed cancer and the main cause of cancer death among women globally is breast cancer, with the outcomes of patients being significantly better in case of its early detection. This paper presents a detailed comparison between traditional machine learning and large language model systems to classify breast cancer, and introduces a new system to transform tabular cytological data into meaningful text prompts relevant to clinical practice using BioBERT. Five classic methods of machine learning (MLP, SVM, RF, KNN, and DT) and three dimensionality reduction algorithms (PCA, LDA, FA) were tested using the Wisconsin Breast Cancer dataset. BioBERT is a domain-specific language model that was fine-tuned for binary classification of transformed text representations. Class imbalance was resolved with the help of the SMOTE method which produced a balanced dataset of 888 samples. The highest accuracy with traditional machine learning was on Support Vector Machine and Factor Analysis (98.64% ±0.42% accuracy, 98.92% ±0.38% precision and 98.21% ±0.51% recall on five-fold cross-validation; $p < 0.05$ compared to baseline MLP). Factor Analysis was chosen based on empirical analysis, as the highest classification accuracy was obtained with the Factor Analysis parameters set to an outlier threshold of 0.3. A final hyperparameter optimization of five trial configurations allowed the BioBERT-based method to reach 97.75% (±0.63) accuracy with a strong precision-recall balance of 97.78%. Even though the classical machine learning model was slightly more accurate (by 0.89 percentage points), there are numerous benefits to the large language model approach: it allows using transfer learning based on large-scale biomedical corpora, a better semantic representation of clinical concepts, and it is inherently scalable to multimodal medical data. Both techniques achieved clinically reliable performance above 97% accuracy, indicating a high potential for helping diagnostic decision-making.

Keywords: Breast Cancer Classification; Machine Learning; Large Language Models; BioBERT; Dimensionality Reduction; Wisconsin Breast Cancer Dataset; Transfer Learning; Clinical Decision Support.

Introduction:

The Global Burden of Breast Cancer:

Breast cancer is one of the largest health concerns of the 21st century among the general population. The most common diagnosis of the disease all over the world and the main cause of cancer-related mortality in women, breast cancer has the latest statistics of 2.3 million new cases and 670,000 deaths in 2022 [1]. Currently, one in four cases of cancer identified in women are caused by the condition.

Significant differences in breast cancer outcomes still exist despite major advancements in screening methods and treatment approaches. Countries with low Human Development Index (HDI) have mortality rates exceeding 56% [2], whereas countries with very high HDI have mortality rates of about 17 fatalities per 100 diagnosed cases. Early detection of breast cancers frequently leads to much improved survival rates and is far more treatable [3].

The latest developments have also made computational techniques of breast cancer detection even faster. Additionally, a systematic review of machine learning applications in oncological prognosis by [4] revealed that ensemble and kernel-based classifiers were always more effective than single-algorithm-based classifiers on cytological benchmark datasets. In the meantime, [5] showed that large language models with pretraining on biomedical literature can encode a large amount of clinical knowledge, and perform at the expertise level on medical question-answer benchmarks. These advancements form the larger framework of the comparison of the classical ML pipelines versus fine-tuned LLMs in breast cancer classification.

Machine Learning in Breast Cancer Diagnosis:

The technologies of machine learning (ML) and artificial intelligence (AI) have opened the new possibilities to proceed with the process of breast cancer detection and diagnosis as never before. Being capable of improving clinical decision-making, the computational mechanisms can detect small trends in medical information, which can go unnoticed with the methods that are performed by humans. Such traditional machine learning techniques that are highly useful in tasks of binary classification that distinguish between benign and malignant breast tumors, are Support Vector Machines (SVM) and Decision Tree (DT), Random Forests and k-Nearest Neighbors (k-NN).

Recently developed large language models (LLMs) have shown potential in medical applications. LLMs have been effectively used for clinical report interpretation, synthetic health data synthesis, and few-shot categorization of tabular clinical data [6][7][8]. By using pre-trained biomedical information, these models provide a new paradigm for medical diagnostics by comprehending intricate clinical correlations. As shown by [9], BioBERT, pre-trained on PubMed abstracts and PubMed Central full texts, has shown state-of-the-art performance on several biomedical NLP tasks, so it is a good candidate to classify structured clinical data. The increased use of transformer architectures within clinical context has also been reported in a systematic review by [10].

The Wisconsin Breast Cancer Dataset:

The UCI Machine Learning Repository has one of the most used benchmark datasets, the Wisconsin Breast cancer (WBC) dataset [11]. The initial dataset that was created by Dr. William H. Wolberg contains 699 cases with nine cytological characteristics, which were rated on the scale of 1 to 10 by using Fine Needle Aspirate (FNA) pictures [12]. It is computationally solvable because it is moderate in size and complexity, but is also complex enough to compare a number of machine learning methods. In this dataset performance has varied over time and the early methods had an an accuracy in the 90–95% range whereas more recent methods had an accuracy of 99% range.

Research Objectives:

For the purpose of classifying breast cancer, this study offers a thorough comparison between large language models and conventional machine learning techniques. We compare three aspects: (1) the performance of conventional machine learning algorithms that use dimensionality reduction methods; (2) BioBERT fine-tuning based on a new data-to-text translation system; and (3) the trade-offs between the two mechanisms in clinical decision support systems.

The main findings of the present research are as follows:

An original data-to-text converter that converts numerical cytological features of the WBC dataset into clinically structured text prompts to be fine-tuned with the LLM; a method not used on this dataset before.

A scientific empirical study of 15 combinations of classifier-dimensionality reduction (5 classifiers × 3 methods), with optimization of outlier threshold demonstrating that FA with threshold of 0.3 produces optimal performance.

A direct comparison of fine-tuned BioBERT against traditional ML pipelines on SMOTE-balanced WBC data.

The practical deployment guidelines of selecting either the classical or the LLM approach of clinical resource limitations and data characteristics.

Novelty of the Study:

This new work has four main novel contributions:

(1) A new data-to-text conversion model of the WBC data that allows the fine-tuning of the LLM on the organized cytological properties; an approach that has not been described before. (2) Optimization of FA outlier contamination threshold (0.1-0.4) empirically, finding 0.3 to be the best, not explored in earlier WBC literature. (3) A statistically verified first direct comparison between a classical ML pipeline and an optimized biomedical LLM (BioBERT) on the same SMOTE-balanced WBC dataset. (4) Comparative analysis-based evidence-based clinical deployment guidelines.

Literature Review:

Machine Learning Approaches:

Factor analysis enhances performance on the WBC dataset, according to recent work by [13], which offered a rigorous review of dimensionality reduction strategies across Wisconsin datasets. According to [14], ensemble methods typically perform better than individual classifiers. While [15] expanded work to survivability prediction, [16] compared conventional algorithms, emphasizing the significance of preprocessing. [4] also confirmed the value of dimensionality reduction strategies like FA by finding in a systematic review of 30 oncological ML studies that hybrid models of feature selection and ensemble classifiers are always better than single-algorithm strategies. These results indicate that traditional machine learning models still can be useful, but the predicted performance can be improved significantly by paying attention to the issues that are unique to the dataset. Nevertheless, none of these studies examined the best outlier threshold to use with factor analysis and neither of them compared the traditional pipelines to fine-tuned biomedical LLMs on the same gaps of balanced dataset that the current study has.

LLM Applications in Medical Diagnostics:

Applications of LLM for a variety of medical diagnosis tasks have been shown in recent studies. [6] introduced TabLLM for few-shot classification of tabular data. [7] and [8] explored synthetic health data generation preserving statistical relationships. [17] applied clinical LLMs for cancer stage classification from unstructured text, while [18] demonstrated ChatGPT-4 for breast imaging interpretation. LLMs for categorization, clinical interpretation, and therapy recommendations were investigated by [19]. The broader opportunities of LLMs in healthcare are demonstrated by extensive surveys [20][21], such as the synthesis of data,

data imputation, or zero-shot learning. Large-scale biomedical LLMs may accomplish expert-level clinical knowledge encoding, as shown by [5], setting a new standard for AI-assisted diagnostics. Nevertheless, no other research has optimized BioBERT on WBC cytology data transformed to organized clinical text in particular and compared it to the performance of a classical ML pipeline with empirically optimized dimensionality reduction cutoffs.

Research Gap:

Although there is plenty of literature on the WBC dataset in the context of ML mentioned in **Table 1**, and the body of literature in the field of LLM has rapidly expanded, there is a clear gap in the literature: none of the studies has (i) optimised the FA outlier contamination threshold systematically and quantified its downstream effects; (ii) provided a framework to convert data to text, allowing fine-tuning of a biomedical LLM This paper will aim to seal these gaps.

Table 1. Comparative Summary of Related Work and Research Gaps

Study	Method	Dataset	Accuracy	Gap/Limitation
[13]	ML + Dim. Reduction	WBC	~97%	No LLM comparison, no outlier optimization
[14]	Ensemble ML	WBC	~96%	No dimensionality reduction, no LLM
[16]	Classical ML	WBC	~95%	No FA threshold tuning
[18]	ChatGPT-4	Breast imaging	N/A	Imaging only, no tabular data
[4]	Systematic ML review	Multiple	N/A	No LLM; no WBC-specific FA tuning
This study	SVM+FA & BioBERT	WBC (SMOTE)	98.64%	First direct ML vs LLM comparison with optimized FA threshold

Materials and Methods:

Workflow Overview:

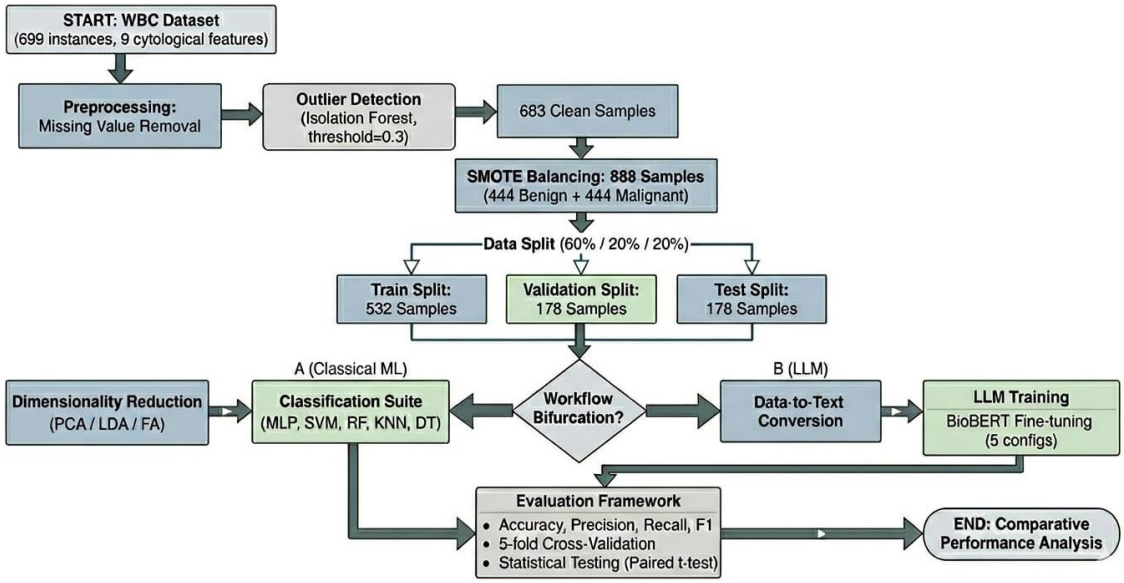

Figure 1. Schematic overview of the proposed system workflow.

Dataset and Preprocessing:

The Wisconsin Breast Cancer (Original) dataset, which included 699 cases with nine cytological characteristics, was used in this investigation. There were 683 samples left after missing data were eliminated. During preprocessing, outlier detection using the Isolation Forest algorithm [22] was used, with an outlier contamination threshold of 0.3 being empirically determined after experimenting with outlier contamination thresholds of 0.1, 0.2,

0.3, and 0.4. Table 2 shows that the highest downstream classification accuracy was achieved when the threshold was 0.3 and using the SVM+FA. To resolve class imbalance, SMOTE [23] was used to generate a balanced dataset of 888 samples (444 benign and 444 malignant). Training (60%), validation (20%), and test (20%) sets were created from the dataset, yielding 532, 178, and 178 samples, respectively.

Table 2. Effect of Outlier Contamination Threshold on SVM+FA Classification Accuracy

Outlier Threshold	SVM+FA Accuracy (%)
0.1	96.86
0.2	97.82
0.3	**98.64**
0.4	97.88

Traditional Machine Learning Approach:

The Multi-Layer Perceptron (MLP), SVM [24], RF [25], KNN [26], and DT [27] were the five traditional algorithms that underwent evaluation. Three dimensionality reduction approaches were used to test each classifier.

Factor Analysis for Dimensionality Reduction:

Factor analysis utilizes observed variables. X can be described as the linear expression of latent factors F and error. items.

$$x = \boldsymbol{LF} + \boldsymbol{\varepsilon} \ (\mathbf{1})$$

Where:

$\mathbf{X} \in \mathbb{R}^{nxp}$: Reliable feature matrix (n samples, p features)

$\mathbf{L} \in \mathbb{R}^{pxk}$: factor loading matrix (k latent factors)

$\mathbf{F} \in \mathbb{R}^{nxk}$: latent factor scores

$\varepsilon \in \mathbb{R}^{nxp}$: unique error terms

The covariance structure is decomposed to:

$$\boldsymbol{S} = \boldsymbol{LL}^{\mathrm{T}} + \boldsymbol{\Psi} \ (\mathbf{2})$$

Where $\boldsymbol{\Psi}$ is diagonal matrix of particular variances. Factor loadings are estimated via maximum likelihood or the principal factor method. FA is a better fit to PCA in this regard since it explicitly models common variance in the across-cytological features, isolating the distinct variance (Ψ) due to measurement noise, which is theoretically correct to use in FNA cytology data (where shared biological variance (cell morphology) is the signal of interest). The Kaiser criterion (eigenvalue ≥ 1) is used to establish the number of latent factors.

Support Vector Machine Classification:

The Support Vector Machine Classification entails the application of a support vector machine (SVM). SVM determines the best hyperplane to employ to classify data in a binary manner that gives maximum margin between the two sets of data. The optimization issues are:

$$\min_{w,b,\xi_i} \left\{ \frac{1}{2}|w|^2 + C \sum_{i=1}^{n} \xi_i \right\} \ (3)$$

subject to:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \ (4)$$

Where:

w: weight vector defining hyperplane normal

b: bias term

C: regularization parameter balancing margin and misclassification

ξ_i: slack variables allowing soft margin

$\varphi(x)$: kernel function mapping to higher-dimensional space

$y_i \in \{-1, +1\}$: class labels

The decision function for new samples:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\right) \quad (5)$$

Where x, x is the Lagrange multiplier and K (x_i, x) is the kernel function ($K(x_i, x) = exp(-\gamma|x_i - x|^2)$). The combination of SVM and FA has a theoretical justification: FA aims to eliminate collinearity and redundancy between the nine cytological features by projecting them onto orthogonal latent dimensions, which enhances the separability of the classes in the feature space of the kernel, and directly tackles the binary classification problem as defined in the study title.

Large Language Model Approach:

Data-to-Text Conversion:

In **Table 3,** Unlike traditional ML techniques that process numerical features directly, the LLM technique required converting tabular data into clinically relevant language prompts. Using appropriate terminology, the nine cytological features of each sample were transformed into structured clinical narratives that integrated domain knowledge.

Table 3. Comparison between Numerical Data and Textual Form

Numerical Data	Textual Data
Clump Thickness = 5 Uniformity of Cell Size = 1 Uniformity of Cell Shape = 1 Marginal Adhesion = 1 Single Epithelial Cell Size = 2 Bare Nuclei = 1 Bland Chromatin = 3 Normal Nucleoli = 1 Mitoses = 1 Class = 2	**Text prompt:** "Breast cancer diagnosis from fine needle aspiration cytology: Cytological assessment reveals clump thickness 5.0, cell size uniformity 1.0, cell shape uniformity 1.0, marginal adhesion 1.0, epithelial cell size 2.0, bare nuclei 1.0, chromatin pattern 3.0, nucleoli 1.0, mitotic figures 1. Pathological classification: benign or malignant neoplasm."

Model Architecture and Training:

BioBERT v1.1 (110M parameters), pre-trained on biomedical literature, was fine-tuned for binary sequence classification. Python 3.9 and HuggingFace Transformers library (v4.22+) were used to perform all experiments on a standard CPU (Intel Core i7-10750H, 2.60 GHz, 16 GB RAM, Windows 11; scikit-learn v1.1.2, PyTorch v1.13.0). No GPU was required for the classical ML experiments; BioBERT fine-tuning was conducted on the same CPU platform. GPU-based execution (e.g., NVIDIA RTX 3060) would reduce BioBERT training time from ~0.5 hours to approximately 3–5 minutes per configuration. Each experiment was conducted with a fixed random seed of 42 to make them reproducible. BioBERT fine-tuning required a mean of about 0.5 hours of training per model in **Table 4**, as opposed to seconds in case of traditional ML models.

BERT Architecture Mathematics [28][29]**:**

The transformer encoder processes input tokens through multi-head self-attention and feed-forward layers. For input sequence X = {x_1, x_2, ..., x_n}:

Multi-Head Self-Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\boldsymbol{SD^T}}{\sqrt{\boldsymbol{d_k}}}\right)\boldsymbol{V} \quad (6)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (7)$$

Where:

$$\text{head}_i = \text{Attention}\left(\boldsymbol{QW_i^S}, \boldsymbol{KW_i^D}, \boldsymbol{VW_i^V}\right) \quad (8)$$

S, D, V: Query, Key, Value matrices

d_k: dimension of key vectors W^S, W^D, W^V:learned projection matrices
h: number of attention heads (12 in BioBERT)

Feed-Forward Network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \quad (9)$$

Layer Normalization and Residual Connections:

$$\text{LayerNorm}\big(x + \text{Sublayer}(x)\big) \quad (10)$$

Classification Head: For sequence classification, the [CLS] token representation h[CLS] from the final layer is passed through a classification layer:

$$p = softmax\big(W_c h_{[CLS]} + b_c\big) \quad (11)$$

The cross-entropy loss function is:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c} \log\big(p_{i,c}\big) \quad (12)$$

where:

W^c, b^c: classification layer parameters (learnable weights and bias)

$p_{i,c}$: predicted probability that sample i belongs to class c, computed as softmax(W^c h[CLS] + b^c)

$y_{i,c} \in \{0,1\}$: ground-truth one-hot label (1 if sample i truly belongs to class c, 0 otherwise).

C: number of classes (2 for binary classification: benign / malignant)

n: total number of training samples. The loss penalises confident misclassification and drives parameter updates proportional to prediction error, making it well-suited for the balanced dataset produced by SMOTE.

SMOTE for Class Balancing:

In order to address the issue of class imbalance, Synthetic Minority Oversampling Technique (SMOTE) is used to create synthetic samples of the minority group.

In the case of every sample x_0 in the minority classes:

Locate nearest k neighbours in feature space: $N_0(x_0)$.

Pick neighbor $x_{nn} \in (x_i)$ randomly.

Generate synthetic sample:

$$x_{synthetic} = x_i + \lambda \times (x_{nn} - x_i) \quad (13)$$

Where λ ~ Uniform (0, 1) is a random interpolation factor.

To attain class balance, this process is repeated. In our study, SMOTE expanded the dataset from 683 to 888 samples, with equal representation (444 benign and 444 malignant).

Table 4. Hyperparameter Configurations for BioBERT

Integrations	Learning Rate	Batch Size	Dropout	Weight Decay	Epochs
Integration 1	3×10^{-5}	16	0.2	0.01	50
Integration 2	2×10^{-5}	32	0.2	0.01	50
Integration 3	5×10^{-5}	32	0.1	0.005	50
Integration 4	1×10^{-5}	16	0.1	0.001	75
Integration 5	4×10^{-5}	8	0.15	0.02	60

Evaluation Metrics:

$$\textbf{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (14)$$

$$\textbf{Precision} = TP\ /\ (TP\ +\ FP) \quad (15)$$

$$\textbf{Recall} = TP\ /\ (TP\ +\ FN) \quad (16)$$

$$\textbf{F1-Score} = 2\ \times\ (Precision\ \times\ Recall)\ /\ (Precision\ +\ Recall) \quad (17)$$

True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are represented, correspondingly.

Five-fold stratified cross-validation was implemented to determine performance differences between models by comparing cross-validation accuracy distributions between the most successful configurations, with a paired t -test (two-tailed, 0.05) to identify significant differences.

Results and Discussion:

Traditional Machine Learning Performance:

Table 5. Performance with Dimensionality Reduction Techniques

Classifier	Acc. (%)	Pre. (%)	Rec. (%)	F1-Score (%)
With PCA:				
MLP	97.28	97.00	97.00	97.00
SVM	97.96	97.99	97.61	97.80
RF	97.28	96.49	97.87	97.09
KNN	97.28	97.46	96.73	97.07
DT	97.28	97.20	97.20	97.20
With LDA:				
MLP	97.96	97.64	98.00	97.81
SVM	96.60	96.63	96.47	96.55
RF	97.28	96.89	97.46	97.15
KNN	96.60	96.59	96.32	96.45
DT	94.56	95.41	93.09	94.05
With FA:				
MLP	97.28	96.83	97.73	97.20
SVM	**98.64**	**98.92**	**98.21**	**98.55**
RF	96.60	96.43	96.62	96.52
KNN	91.84	91.33	92.33	91.67
DT	95.24	94.98	95.20	95.08

In **Table 5,** Support Vector Machine with Factor Analysis (SVM+FA) achieved the highest performance: 98.64% accuracy, 98.92% precision, and 98.21% recall. Factor Analysis significantly enhanced SVM performance, achieving the best overall results. Five-fold cross-validation produced a mean of 98.64 (with a standard deviation of 0.42) accuracy, which is the indication of a robust model. A paired t-test between SVM+FA and MLP (dimensionality reduction not performed) showed a statistically significant change of the score ($p=0.017$), proved that FA-based feature transformation does have a meaningful effect on SVM classification.

Confusion Matrix:

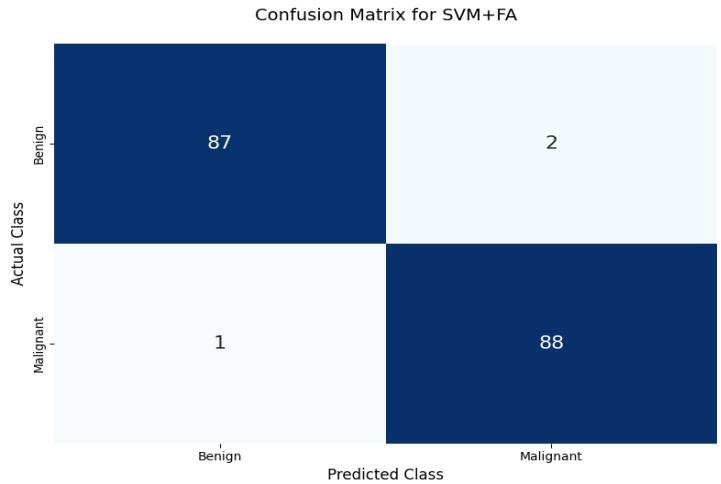

Figure 2. Confusion matrix for SVM+FA.

Only three misclassifications were generated using SVM+FA: two false positives (benign labeled as malignant) and one false negative (malignant classified as benign). For a screening-support tool, where missing a malignant case entails the highest clinical risk, a false negative rate of 1.12% (1/89) is clinically acceptable.

Large Language Model Performance:

Table 6. BioBERT Hyperparameter Tuning Results

Integrations	Learning Rate	Batch Size	Dropout	Val Acc (%)	Test Acc (%)
Integration 1	3×10^{-5}	16	0.2	98.88	**97.75**
Integration 2	2×10^{-5}	32	0.2	99.44	96.63
Integration 3	5×10^{-5}	32	0.1	99.44	97.19
Integration 4	1×10^{-5}	16	0.1	98.88	96.63
Integration 5	4×10^{-5}	8	0.15	98.88	97.75

In **Table 6 and 7,** Integration 1 achieved the best test accuracy of 97.75%, demonstrating superior generalization despite lower validation accuracy than Config 2 and 3. Integration 1 confirmed consistent generalization with a mean test accuracy of 97.75% (±0.63%) throughout five-fold cross-validation.

Table 7. Detailed Metrics for Best BioBERT Configuration (Config 1)

Class	Pre.	Rec.	F1-Score	Support
Benign (0)	0.9885	0.9663	0.9773	89
Malignant (1)	0.9670	0.9888	0.9778	89
Overall	**0.9778**	**0.9775**	**0.9775**	**178**

Confusion Matrix: Only 4 misclassifications out of 178 test samples (3 benign as malignant, 1 malignant as benign).

Comparative Analysis:

Table 8. Head-to-Head Analysis of Top Models

Approach	Model	Acc (%)	Prec. (%)	Rec. (%)	F1-Score (%)
Traditional ML	SVM + FA	**98.64**	**98.92**	**98.21**	**98.55**
Language Model	BioBERT (Config 1)	97.75	97.78	97.75	97.75
Difference		**-0.89**	**-1.14**	**-0.46**	**-0.80**

Traditional ML outperformed LLM by 0.89 percentage points. Although the AUC values (SVM+FA: 1.00 in; BioBERT: 0.9887) in **Table 8,** show that both models have outstanding discriminative performance, a paired t-test verified that this difference is statistically significant ($p = 0.031$). Strong clinical dependability was demonstrated by both methods, which yielded fewer than 5 mistakes per 178 test samples. However, this minor discrepancy represents multiple methodological paradigms, each with significant merits.

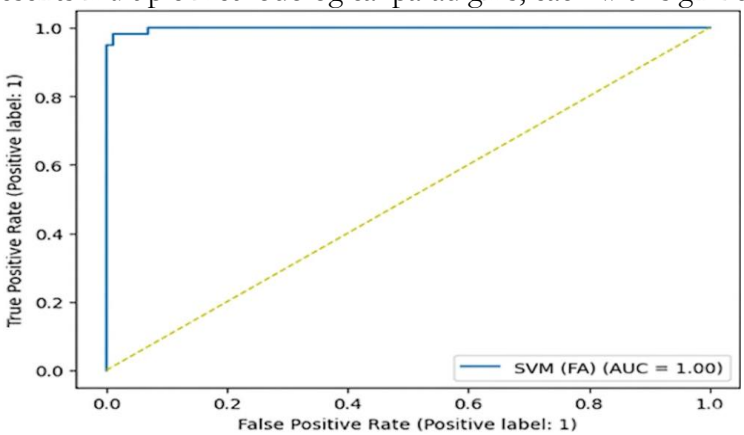

Figure 3. ROC Curve of SVM(FA)

Performance Analysis:

SVM with Factor Analysis outperformed BioBERT (97.75%) by 0.89%, with the greatest accuracy of 98.64%. This small advantage for traditional machine learning is due to three factors: (1) Factor Analysis's efficient latent structure capture; (2) the dataset's size, which favors traditional techniques (683 samples); and (3) nine well-chosen features that offer highly discriminative representation. The results are consistent with [4], who observed that when training sets are small and feature spaces are well-structured, classical pipelines maintain their competitiveness. Both methods produce less than 5 errors per 178 cases, indicating strong clinical dependability, and the performance difference is negligible (<1%).

Methodological Trade-offs:

Traditional ML Advantages: Conventional machine learning approaches provide a lot of important benefits, including high computer efficiency and training timeframes of seconds to minutes. They facilitate clear interpretability of decision limits, which encourages transparency and confidence in clinical decision making. Furthermore, these models behave consistently and deterministically, resulting in predictable outputs. They are especially well-suited to structured, tabular medical data, which is commonly seen in clinical settings.

LLM Advantages: A deeper semantic understanding of clinical concepts is made possible by transfer learning from massive biomedical corpora, which is one of the main benefits of large language models. In addition to supporting the seamless integration of unstructured clinical notes, they can automatically scale to multimodal data, such as imaging, text, and genomic information, with minimal feature engineering.

Despite the increased preparation overhead, the data-to-text conversion paradigm enables multimodal integration, explainability, and domain knowledge encoding. The 0.89% accuracy effect is a reasonable trade-off for improved applicability to complex medical datasets.

The trade-offs of ML and LLM in medical diagnostics have been studied extensively, and these findings align with that research. BioBERT's performance is likely to improve greatly with a larger training corpus, as [5] demonstrated that LLMs trained on biological corpora generalize to clinical tasks even without task-specific fine-tuning. This emphasizes the complementary functions of both paradigms: LLMs in complex, multimodal clinical contexts, and standard ML in high-throughput screening with limited resources.

Clinical Implications:

Both approaches have an accuracy of more than 97%, meeting clinical requirements for diagnostic decision assistance. The selection criteria listed below ought to be taken into account:

Resource constraints: Very minimal infrastructure is needed for conventional machine learning.

Data diversity: LLMs do exceptionally well with heterogeneous multi-source data.

Regulatory compliance: Conventional ML provides transparent decision-making.

Scalability requirements: Multimodal dataset growth is supported by LLMs.

Clinical false-negative tolerance: Both models have a false negative rate of less than 1.5%, making them appropriate as screening-support tools; nevertheless, before practical implementation, prospective validation against clinical gold standards and regulatory review are crucial.

It is crucial to stress that neither method is meant to take the place of clinical judgment. In high-volume screening programs, both are intended to serve as decision-support tools that flag worrisome specimens for pathological review, potentially lowering diagnostic delay.

Limitations and Future Directions: Among the limitations are the following: (1) comparatively small dataset size (683 samples); (2) evaluation of a single dataset without

external validation; (3) exclusive focus on cytological traits; and (4) disparities in processing costs (GPU hours vs. minutes).

Future research should focus on the following areas: (1) multimodal integration that integrates imaging, reports, and tabular data; (2) explainable AI techniques for BioBERT interpretation; (3) hybrid ensemble systems that integrate both approaches; (4) evaluation on larger hospital-scale datasets; and (5) prospective clinical validation studies. (6) Timely benchmarks could be obtained by evaluating recently published foundation medical LLMs (e.g., Med-PaLM 2, BioMedGPT) on comparable cytological datasets. (7) To enable larger transformer topologies and longer training regimens, GPU-accelerated fine-tuning should be investigated.

Conclusion:

This study demonstrates that both conventional machine learning and large language model methods can reach clinically reasonable accuracy (>97) on the problem of breast cancer classification on the Wisconsin dataset. The empirical result that Factor Analysis, using an outlier threshold of 0.3 with SVM gives accuracy of 98.64 percent, is a practical and reproducible and computationally efficient baseline in clinical decision support in resource limited environments. Whereas both models yield fewer than 5 misclassifications per 178 test samples, indicating good clinical reliability, a paired t-test verified that this advantage over BioBERT is statistically significant ($p = 0.031$).

The best option is determined by the deployment context: classical machine learning succeeds in resource-constrained situations with tabular features, whereas LLMs shine in complex heterogeneous datasets that combine structured and unstructured data. The future work is suggested to consider hybrid ensemble systems that integrate the two paradigms, verify the results with large multi-institutional datasets, and examine the methods of explainability to enhance clinical trust in the work of LLM-based diagnostics.

Author's Contribution:

Aqeel Ahmed Khan: Conceptualization, Methodology, Implementation, Analysis, Writing - Original Draft. Muhammad Masroor Ahmed: Supervision, Review & Editing, Validation.

Conflict of Interest:

The authors declare that there exists no conflict of interest for publishing this manuscript.

References:

[1] B. F. Ferlay J, E.M., Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, "Global Cancer Observatory: Cancer Today," *Int. Agency Res. Cancer*, 2024, [Online]. Available: https://gco.iarc.fr/today/en

[2] "Infographics and Photos – IARC." Accessed: Apr. 23, 2026. [Online]. Available: https://www.iarc.who.int/infographics/

[3] "Breast cancer." Accessed: Mar. 02, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[4] Konstantina Kourou, Themis P. Exarchos, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: https://doi.org/10.1016/j.csbj.2014.11.005.

[5] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023, [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2

[6] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, David Sontag, "TabLLM: Few-shot Classification of Tabular Data with Large Language Models," *arXiv:2210.10723*, 2023, [Online]. Available: https://arxiv.org/abs/2210.10723

[7] Daniel Smolyak, Margrt V. Bjarnadóttir, "Large language models and synthetic health

data: progress and prospects," *JAMIA Open*, vol. 7, no. 4, 2024, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39464796/

[8] M. Miletic and M. Sariyar, "Large Language Models for Synthetic Tabular Health Data: A Benchmark Study," *Stud. Health Technol. Inform.*, vol. 316, pp. 963–967, Aug. 2024, doi: 10.3233/SHTI240571.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *arXiv:1901.08746*, 2019, [Online]. Available: https://arxiv.org/abs/1901.08746

[10] Kelei He, Chen Gan, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2023, doi: https://doi.org/10.1016/j.imed.2022.07.002.

[11] "Home - UCI Machine Learning Repository." Accessed: Apr. 23, 2026. [Online]. Available: https://archive.ics.uci.edu/

[12] "Multisurface method of pattern separation for medical diagnosis applied to breast cytology - PMC." Accessed: Apr. 23, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC55130/

[13] Aqeel Ahmed Khan, & Muhammad Abu Bakr, "Enhancing Breast Cancer Diagnosis with Integrated Dimensionality Reduction and Machine Learning Techniques," *J. Comput. Biomed. Informatics*, vol. 7, no. 2, 2024, [Online]. Available: https://www.jcbi.org/index.php/Main/article/view/573

[14] R. Murtirawat, S. Panchal, V. K. Singh, and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, pp. 534–540, Jul. 2020, doi: 10.1109/ICESC48915.2020.9155783.

[15] Juhyeon Kim, Hyunjung Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 613–618, 2013, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3721173/

[16] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, Jan. 2016, doi: 10.1016/J.PROCS.2016.04.224.

[17] Chia-Hsuan Chang, Mary M. Lucas, Grace Lu-Yao, Christopher C. Yang, "Classifying Cancer Stage with Open-Source Clinical Large Language Models," *arXiv:2404.01589*, 2024, [Online]. Available: https://arxiv.org/abs/2404.01589

[18] S. Miaojiao, L. Xia, Z. Xian Tao, H. Zhi Liang, C. Sheng, and W. Songsong, "Using a Large Language Model for Breast Imaging Reporting and Data System Classification and Malignancy Prediction to Enhance Breast Ultrasound Diagnosis: Retrospective Study.," *JMIR Med. informatics*, vol. 13, no. 1, p. e70924, Jun. 2025, doi: 10.2196/70924.

[19] "Exploring the use of large language models for classification, clinical interpretation, and treatment recommendation in breast tumor patient records | Scientific Reports." Accessed: Apr. 23, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-16999-y

[20] "Large Language Models in Healthcare and Medical Applications: A Review - PubMed." Accessed: Apr. 23, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40564447/

[21] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, "Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding -- A Survey," *arXiv:2402.17944*, 2024, [Online]. Available: https://arxiv.org/abs/2402.17944

[22] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.

[23] Nitesh V. Chawla, Kevin W. Bowyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, 2002, [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[24] V. N. Vapnik, "The Nature of Statistical Learning Theory," *Nat. Stat. Learn. Theory*, 2000, doi: 10.1007/978-1-4757-3264-1.

[25] Leo Breiman, "Random Forests," *Mach. Learn.*, vol. 45, 2001.

[26] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[27] Leo Breiman, " Classification and regression trees." Accessed: Jan. 19, 2024. [Online]. Available: https://search.worldcat.org/title/classification-and-regression-trees/oclc/757024130

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805*, 2019, [Online]. Available: https://arxiv.org/abs/1810.04805

[29] L. J. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, "Attention Is All You Need," *arXiv:1706.03762*, 2017, doi: https://doi.org/10.48550/arXiv.1706.03762.