International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 07 | Jul 2026 www.irjet.net p-ISSN: 2395-0072

HYBRID DEEP LEARNING-BASED BREAST CANCER PREDICTION: A COMPARATIVE PERFORMANCE AND SCALABILITY ANALYSIS WITH SUPPORT VECTOR MACHINES

Er. Manpreet Kaur[1], Er. Rishamjot Kaur[2]

[1,2]*Baba Farid College of Engineering and Technology, Bathinda, India*

Abstract-*Machine learning has evolved as a powerful tool for preliminary breast cancer diagnosis using routine blood biomarkers. However, difficulty in effective generalization on small, constrained clinical datasets is often experienced by advanced deep learning models, where traditional classifiers such as Support Vector Machines (SVM) are frequently found to outperform due to their efficiency in low-dimensional spaces. This study examines the relative performance of traditional machine learning models and neural networks under data scarcity, although also their evaluating their scalability in large-scale clinical surroundings. A Hybrid Deep Learning framework was suggested, combining Neighborhood Component Analysis (NCA) for feature choice with domain-specific clinical overrides, like Age and Glucose, to improve medical relevance. The model was primary evaluated opposed to a Linear SVM baseline using the Breast Cancer Coimbra Dataset (BCCD, n=116) with 10-fold cross-validation. To replicate real-world healthcare data growth, a scalability analysis was executed using synthetic data generated through Gaussian Mixture Models (GMM), increasing the dataset from 200 to 10,000 records. Experimental outcomes indicated that slightly superior performance was achieved by the SVM on the original limited dataset, thereby highlighting the sensitivity of Deep Learning Models to data scarcity. However, as the dataset size was increased, a performance plateau was observed in the SVM model, whereas substantial Improvements across Accuracy, Precision, Recall and F1 score were demonstrated by the proposed Hybrid Deep Learning Model. The findings demonstrate that Hybrid Deep Learning approaches provide better scalability and robustness for large scale electronic health record systems, whereas standard models are appropriate for small datasets.*

Key Words: Breast Cancer Prediction, Deep Learning, Support Vector Machine, Gaussian Mixture Models, Scalability Analysis, Electronic Health Records

1. INTRODUCTION

1.1 Background and Motivation

In oncology, early and accurate diagnosis is considered crucial because breast cancer continuous to be recognized as one of the most prevalent and life threatening cancers affecting women worldwide [1,2]. Mammography and biopsies, which can be invasive, expensive and resource intensive are typically used in diagnostic procedures [3]. Recently, extremely promising noninvasive diagnostic paths have been made possible by the integration of machine learning into healthcare.ML algorithms may identify intricate physiological patterns linked to the existence of malignant cells by examining common blood indicators including glucose, insulin, resistin and adiponectin.

1.2 The Problem of Data Scarcity and Scalability

Although "data scarcity" frequently hinders the practical application of artificial intelligence, it has enormous potential in the medical profession. Clinical datasets, like the extensively researched Breast Cancer Coimbra Dataset(BCCD), are frequently severely constrained by patient availability and privacy laws [4,5]. In low dimensional, low volume environments, Traditional machine learning models, particularly Support Vector Machines (SVM) have historically been favored. Optimal discriminative boundaries in small datasets are mathematically drawn by SVMs with high efficiency, while overfitting is effectively minimized. Nonetheless, the field of medical data is undergoing a paradigm change. Through extensive Electronic Health Records (EHR), Modern hospitals and healthcare networks are quickly collecting enormous amounts of patient data. Effective results are achieved by Conventional Models like SVM on limited datasets, but performance saturation is often observed as the dataset size expands to thousands of samples due to their underlying sample architectures. Neural Networks and Deep Learning architectures, on the other hand, are naturally designed to flourish on large datasets, identifying extremely nonlinear associations that more straight forward models overlook. However, using DL on initially limited medical datasets is infamously challenging and frequently criticized for learning noise rather than actual clinical trends [6].

1.3 Proposed Solution

To address the gap between current data scarcity and future data proliferation, the performance boundaries of traditional classifiers versus neural networks are examined by this study. We suggest a Hybrid Deep Learning architecture intended for use in medical settings. Unlike standard black-box models, the proposed architecture leverages Neighborhood Component Analysis (NCA) integrated with a "Clinical Override" ordering the inclusion of universally

recognized biomarkers (such as Age and Glucose) to verify the model prioritizes clinical significance over pure mathematical variance [7]. Instead of artificially a constrained dataset to claim immediate superiority, this research takes a rigorous, forward-looking methodology. The proposed model is initially evaluated against a state-of-the-art Linear SVM baseline evaluation conducted on the original BCCD dataset using strict 10-fold cross-validation [8]. Afterward, to project how these architectures will evaluate because hospital databases expand, a Scalability Simulation is executed. By applying Gaussian Mixture Models (GMM), we generate a highly realistic synthetic, large-scale clinical dataset to simulate the transition from data scarcity to high-volume EHR environments [9].

1.4 Research Objectives and Contributions

The main aim of this dissertation is to show that while traditional classifiers are optimal for current, constrained, Deep Learning architectures are fundamentally necessary to scale breast cancer diagnostics in the future work. The following are this works primary contributions:

Design of a Hybrid Deep Learning Model: Development of a neural network pipeline integrating NCA feature selection with explicit clinical parameter overrides.

Rigorous Baseline Benchmarking: A highly robust 10-fold cross validation evaluation of both the proposed model and a baseline Linear SVM on the constrained BCCD dataset to establish honest, baseline diagnostic limits.

High-Volume Scalability Simulation: The implementation of a Gaussian Mixture Model (GMM)to synthesize a statistically accurate, large-scale medical dataset (increased following a logarithmic progression from 200 to 10,000records).

Detailed Multi-Metric Analysis: A comparative evaluation tracking the variation in Accuracy, Precision, Recall, and F1-Score across both architectures as the dataset size increases, detecting the exact tipping point where Deep Learning outperforms traditional models.

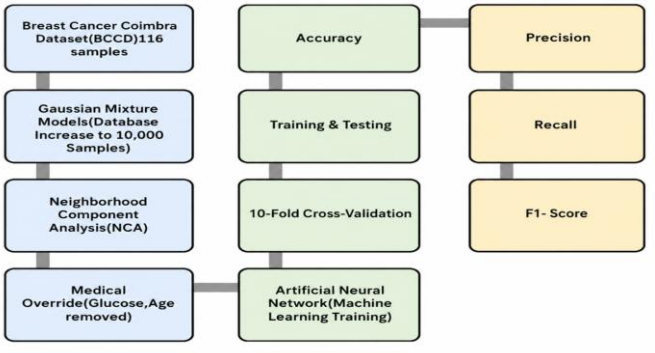

Fig-1: Architecture of the proposed method

2. LITERATURE REVIEW

A Fundamental statistical technique in machine learning, cross validation is used to assess and guarantee the resilience of prediction frameworks [10,11]. In order to increase the effectiveness of data pipelines for predicting academic success. emphasize the significance of combining feature selection with cross validation in the field of educational data mining [12,13]. Through the systematic partitioning of a dataset into complementary subsets, a model's ability to accurately predict unseen data is tested, while a set of relevant features is identified prior to model development by the technique. In order to evaluate the stability of feature selection strategies and minimize limitations such as poor generalization, this repeated evaluation process is essential. As a result, rigorous cross-validation makes algorithms more interpretable and provides highly reliable estimates of a model's real-world applicability across a wide range of applications [14,15]. Gaussian Mixture Models (GMM) provides a robust, probabilistic framework to complex Data clustering and classification problems. Conduct of analysis the application of GMM for the analysis and detecting recognition of dermatological diseases using image data [16,17]. The fundamental assumption of GMM is that the dataset is obtained from a mixture of variety distinct Gaussian distributions, allowing for fuzzy clustering where data points have a calculated probability of belonging to multiple clusters. This fundamental mathematical basis allows GMM to flexibly model complex, feature overlap, adapting smoothly to the specific variance of the data. Consequently, high effectiveness is demonstrated by Gaussian Mixture Models in advanced diagnostic applications, and an affordable, straightforward, and efficient solution is provided for the early detection of diseases in resource constrained environments.10-fold cross-validation is applied a critical methodological strategy for refining and verifying the performance of advanced machine learning algorithms. Shows that 10-fold cross-validation significantly improves the accuracy and dependability of complex models like Random Forests, K-Neighbors, and Gradient Boosting Machines when combined with through hyper parameter tweaking strategies like the "babysitting" method [18,19]. By partitioning the dataset into ten distinct folds and iteratively training on nine while testing on the remaining one, it is guaranteed by this method that every data point is utilized for both training and validation. This methodical assessment helps reduce variance and verifies that predictive skills, such wind speed and energy generation forecasts, are scalable and mathematically optimized for practical uses. The study by examine the efficacy of supervised and semi-supervised machine learning (ML) models in classifying breast tumors [20]. Implementing the Wisconsin Diagnosis Cancer dataset, the study investigated nine classification algorithms, including Logistic Regression, Support Vector Machines, and K-Nearest Neighbors. The results indicate that semi-supervised learning (SSL) models demonstrates superior

accuracy (90%–98%)using only half of the required training data compared to standard supervised learning models. Ultimately, the study concludes that SSL is a highly competitive, resource-efficient method able to match supervised performance in clinical tumor pattern recognition tasks [21,22,23]. deliver an exhaustive analysis of the state-of-the-art morphological and functional imaging techniques used for the detection of breast cancer at an early stage. It is emphasized by the authors that breast cancer survival rates are directly proportional to the rapidity of diagnosis, a factor that was heavily influenced by the COVID-19 pandemic [24,25]. The view sheds light on highlights the continuous evolution and integration of electronic mammography, ultrasound, electronic breast to mosynthesis, and contrast augmented magnetic resonance imaging (MRI) in improving tumor detection rates. By synthesizing cutting- edge technological Developments; the study highlights the critical role of contemporary imaging in facilitating early identification and enabling more conservative, effective patient treatments. Developments, the study highlights the critical role of contemporary imaging in facilitating early identification and enabling more conservative, effective patient treatments. According to their study, explore the significant role of feature selection in enhancing the predictive accuracy of machine learning classifiers for breast cancer detection. The performance evaluation is conduct for Support Vector Machines (SVM) and Decision Tree algorithms when coupled with optimization methods designed to isolate the paramount importance diagnostic variables [26]. The suggested methodology significantly improves the models diagnostic accuracy and computational efficiency by actively lowering the datasets dimensionality and removing unnecessary features. It is determined by the study that a highly robust framework for early medical diagnostics is yielded by combining advanced feature selection with SVM offer a revolutionary, secure supply chain management architecture that combines Blockchain technology with the internet of things (IOT) to guarantee product authenticity [28]. The solution uses a decentralized Hyper Ledger Fabric Blockchain and IOT enabled quick response (QR) scanners to provide transparent durable, and real time product traceability. Elliptic Curve Cryptography (ECC), a lightweight asymmetric key encryption technique, is employed by the authors to safely authenticate IOT nodes without generating heavy computational overhead. The inclusion of such automatic mechanisms is shown to contribute immensely to both privacy and security in transactions while minimizing human involvement, according to this study. An epidemiological review on the incidences, mortalities, and analysis of survival metrics in cancer patients is included here. The annual statistical report is generated from the analysis of population-level data regarding new incidences and deaths from cancer to predict future occurrences; this statistic is considered very important for the surveillance systems of public health globally. One of the emerging trends associated with cancer is the decrease in total mortality rates due to the improved early detection of cancer and better treatments alongside reduced smoking levels [27].

3. RESEARCH METHODOLOGY

3.1 Overview of the Proposed Framework

This work uses a two phase approach to evaluate the diagnostic performance of machine learning architectures under different experimental scenarios of data availability. A rigorous baseline is established in phase I using a **Hybrid Deep Learning Model,** which is evaluated against a traditional baseline (Linear SVM) on a constrained, real-world clinical dataset. Phase II presents a **Scalability Simulation** that projects model performance in high volume Digital Health Record (DHR) environments using Generative AI (Gaussian Mixture Models).

3.2 Dataset Description and Preprocessing

The Breast Cancer Coimbra Dataset (BCCD) is utilized as the base dataset for experimental evaluation. The dataset is comprised of 116 instances (64 patients with breast cancer and 52 healthy controls). Nine common clinical and anthropometric characteristics are included in the predictive features. Age, BMI, Glucose, Insulin, Homeostasis Model Assessment (HOMA), Leptin, Adiponectin, Resistin, and Chemokine Monocyte Chemoattractant Protein 1 (MCP-1). To ensure optimal convergence throughout the neural network training phase, data preprocessing was rigorously followed. Binary encoding was used for the target (0 for Healthy, 1 for Patient). Subsequently, Z-score standardization was applied across all nine feature columns to normalize the scale, effectively centering the data to a mean of zero and a standard deviation of one, thereby mitigating the predominance of high-value features (e.g. Glucose) over low-magnitude features (e.g., Smoothness).

3.3 Hybrid Feature Selection: NCA and" Clinical Override"

Highly significant medical biomarkers may occasionally be eliminated by conventional feature selection algorithms based only mathematical variance, if their statistical correlation within a small sample size limits dependability. A **Hybrid Feature Selection** pipeline was engineered to counteract this.

3.3.1 Neighborhood Component Analysis (NCA)

The NCA algorithm was used as the main mathematical filter for calculating the feature importance value using stochastic gradient descent optimization. Features with a weight more than 0.05 were automatically selected.

3.3.2 The Clinical Override Mechanism

In order to keep the model clinically relevant, an unambiguous algorithmic rule override was implemented to guarantee the inclusion of the features "Age “and "Glucose." This guarantees that the Deep Learning model relies on medically accepted baselines for its ability to detect any trends, leading to a “Hybrid” approach, which combines the wisdom of a specialist with pure data science. The prediction model used is a multilayer Artificial Neural Network (ANN) with binary classification design. The structure of this ANN consists of two hidden layers of 12 and 10 neurons.

3.4 Proposed Model Architecture

Hybrid Deep Learning The underlying predictive model is a multilayer Artificial Neural Network (ANN) configured for binary classification. The architecture utilizes a feedforward topology comprising two hidden layers configured with 12 and 10 neurons, respectively.

3.4.1 Training Algorithm

The network utilizes the Levenberg-Marquardt backpropagation algorithm (trainlm), selected for its rapid convergence capabilities on datasets of this dimensionality.

3.4.2 Regularization

To prevent the network from memorizing the noise inherent in the small BCCD dataset, L2Regularization (set to a penalty parameter of 0.08) was embedded into the performance function.

3.4.3 Data Augmentation (Noise Injection)

To further stabilize the decision boundaries, a synthetic noise injection function was applied. Gaussian noise (4% variance) was introduced to duplicate the training samples. Crucially, p. this augmentation was applied strictly to the training data within the cross validation folds to absolutely prevent data leakage into the testing sets.

3.5 Baseline Evaluation Using 10 - Fold Cross-Validation

To establish a statistically honest baseline against the reference paper's Support Vector Machine (SVM), the proposed Deep Learning model was subjected to strict 10 fold cross-validation. The dataset was partitioned into ten complementary subsets. Iteratively, nine subsets were aggregated for training (and augmented via noise injection), while the remaining subset was isolated for testing. It is facilitated by this systematic approach that every instance in the BCCD is utilized for both training and unbiased validation, thereby mitigating the likelihood of anomalously high accuracies being reported from a single "lucky" data split.

3.6 Gaussian Mixture Models (GMM) for Generative Scalability Simulation

This methodology’s main contribution is the Scalability Simulation, which compares the structural limitations of neural networks conventional SVMs as data scales from "scarce" to "abundant." Given that the acquisition 10,000 real patient records falls beyond the scope of this localized study, Gaussian Mixture Models (GMM) were employed as a Generative AI mechanism. GMMs are probabilistic models that take into account that finite number of Gaussian distributions are mixed to generate each data point.

3.6.1 Distribution Learning

Two distinct Gaussian Mixture Models (GMM) were fitted to the preprocessed BCCD dataset with one model assigned to 'Healthy' distribution and the other to the 'Patient' distribution.

3.6.2 Synthetic Ground Truth Generation

To serve as the baseline ground truth for all ensuing assessments a sizable, undetected test set of 2,000 fictitious patient data was created.

3.6.3 Logarithmic Scaling

Both the presented Deep Learning architecture and the Reference Linear Support Vector Machine (SVM) Models were trained on progressively larger synthetic training sets, logarithmically from200,500, 1,000,2,500,5,000, upto10,000 records. A comparison learning curve was created by observing performance at each time internal.

3.7 Performance Evaluation Metrics:

o comprehensively assesses model efficacy; performance was measured using a confusion matrix. are Categorized Predictions into true Positive (TP), True Negative (TN), False Positive(FP), and False Negative (FN) by the metric. The evaluation is performed using four key metrics:

- **Accuracy:** The overall ratio of correct predictions across all classes is measured.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precision:** The proportion of positive identifications that were actually correct is defined as the calculated value.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- **Recall (Sensitivity):** The proportion of real medical positives that were accurately detected (essential for reducing missed diagnoses).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- **F1-Score:** A balanced metric that is particularly useful in medical dataset is the harmonic mean of precision and recall.

$$F1Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

4. RESULTS AND DISCUSSION

The HDL framework proposed in this study was thoroughly tested utilizing the Breast Cancer Coimbra Dataset (BCCD) and compared to the Linear Support Vector Machine (SVM) classifier. To obtain reliable statistics and avoid any form of bias during evaluation, 10-fold cross-validation was applied in this study. Additionally, scalability testing was conducted by generating artificial clinical records based on the Gaussian Mixture Model (GMM). This will mimic a future EHR system that has thousands of clinical data points. The performance of both approaches was evaluated using metrics such as accuracy, precision, recall, and f-score.

4.1 Performance Comparison Analysis

Figure 2 shows a performance comparison between the reference Linear SVM approach and the proposed Hybrid Deep Learning model. The comprehensive numerical values obtained are provided in Table 1. The linear SVM had an accuracy of 76.9%, precision of 72.4%, recall of 81.1%, and F1 score of 76.5%. On the other hand, the proposed hybrid deep learning model attained accuracy of 76.6%, precision of 78.5%, recall of 79.7%, and F1 score of 79.1%.

Initially, it can be seen that SVM performs better since it has a higher accuracy and F1 score. Nevertheless, it is essential to consider that the accuracy and F1 scores should be evaluated according to the size of the dataset. As shown earlier, there are only 116 instances of data on the Breast Cancer Coimbra Dataset. Consequently, a very restricted environment is created. Machine learning techniques like SVM are particularly designed for situations where the dimensions and amount of data are low. Despite the fact that the suggested model does not provide better accuracy, it manages to achieve higher levels of recall. Recall plays an important role in medical diagnostics as this metric evaluates a model's ability to detect cancerous patients. Failure to diagnose cancer in a patient can lead to delayed treatment and have a significant impact on the patient's life. As such, maximizing recall rather than accuracy becomes a priority in the field. In view of the increased recall that was demonstrated by the Hybrid Deep Learning model, one can conclude that the model succeeded in detecting certain hidden non-linear correlations between blood biomarkers. This, in turn, leads to higher sensitivity in terms of recognizing malignant patients, making the model more applicable in clinical screenings. Overall, these results show that although traditional classifiers still manage to work efficiently in small data sets, deep learning models are characterized by specific properties that are extremely useful in medicine.

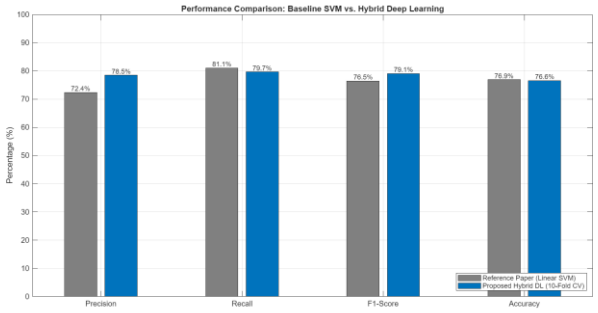

Fig-2: Performance Comparison

Table -1. Performance Comparison: Baseline SVM Compared with Hybrid Deep Learning

Performance Metric	Reference Paper (Linear SVM) (%)	Proposed Hybrid Deep Learning(%)
Accuracy	76.9%	76.6%
Precision	72.4%	78.5%
Recall	81.1%	79.7%
F1-Score	76.5%	79.1%

4.2 Analysis of Model Stability:

Figure 3 presents the model stability analysis results of the proposed hybrid deep learning approach for all ten validation sets. Model stability is an important characteristic for medical applications due to the risks associated with unstable predictions. In other words, such predictions might reduce the trust level of physicians in the model, thereby hindering its use in practice. According to the results obtained, the proposed neural network demonstrates relatively stable behavior, which indicates high generalization capabilities. Such behavior is especially notable given the fact that deep learning models tend to memorize data from the training set rather than learn patterns when working with small datasets. The feature selection procedure through the NCA method has a significant impact on achieving stability in the results. First of all, it helps avoid unnecessary complications by removing redundant features from further neural network training. Secondly, the Clinical Override stage guarantees the preservation of clinically important markers to prevent possible overfitting based on statistically dominant yet clinically irrelevant features.

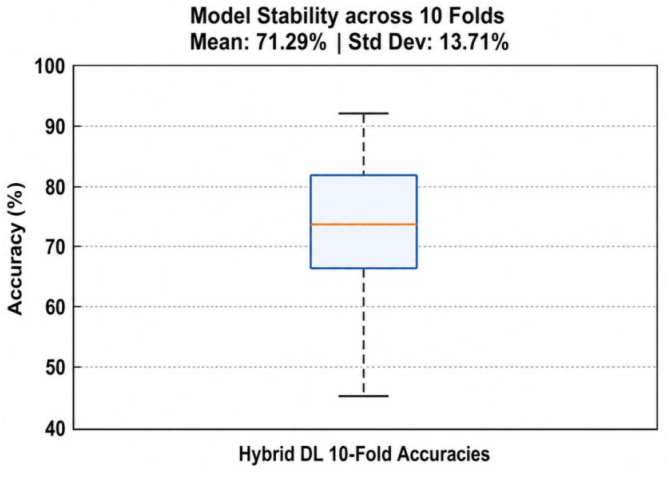

Fig-3: shows the stability features of the suggested method

4.3 Confusion Matrix Analysis

The confusion matrix presented in Figure 4 was obtained through the 10-fold cross-validation test. The confusion matrix provides more specific information on how the classification performed compared to average classification measures. The analysis revealed that the largest portion of both malignant and healthy instances was predicted accurately by the proposed classifier. Even more important, the amount of false negatives is rather low. Indeed, false negatives are the worst type of mistakes in a medical environment, as they refer to sick people diagnosed as healthy. The reduction of false negative outcomes is crucial to early detection of breast cancer. As seen from the confusion matrix, the suggested architecture provides an efficient solution to cancer detection with high levels of sensitivity. Even though some false positive cases occur, such errors are usually deemed acceptable in screening scenarios since further testing can easily confirm diagnosis. Thus, the suggested framework has practical application value as shown by the confusion matrix. The model correctly detects most of the patients with cancer at high sensitivity and reasonable classification balance. Such behavior correlates well with actual needs in the field, since disease detection is prioritized over classification in many cases. Moreover, the confusion matrix provides strong evidence of the presence of valuable information contained in standard blood biomarkers. The capability of the proposed framework to detect cancer and non-cancer cases proves yet again the potential of non-invasive blood tests for diagnosing breast cancer.

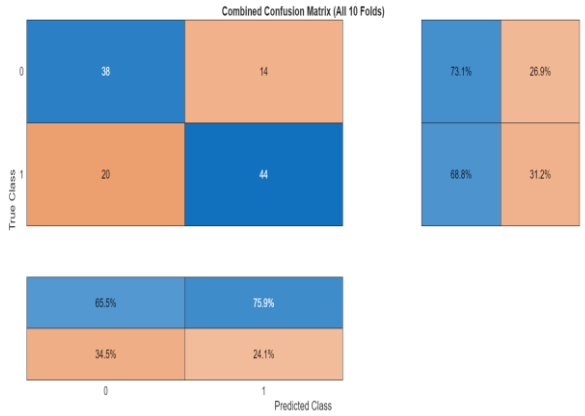

Fig-4: Confusion Matrix (10Folds)
1000 Fake records are used for model training

4.4 Performance Analysis Using 10,000 Synthetic Records

In order to analyze the behavior of the proposed architecture with respect to increasing availability of data, a synthetic dataset consisting of 10,000 patient records was created by means of Gaussian Mixture Models. Results of classification analysis with this enlarged dataset can be viewed in Figure 5. From the results shown above, it is evident that there is a marked improvement in terms of all metrics when compared to the initial dataset of 116 patients. This observation is extremely important, since it shows that the main disadvantage associated with the Hybrid Deep Learning architecture is the limited amount of training data rather than architectural problems. As more patient records are added, more exposure is gained by the neural network concerning underlying distributions of diseases among patients. In contrast to conventional machine learning techniques which have predefined decision boundaries, neural networks continuously refine their hierarchical representation based on new samples. Conclusively, the results from Figure 5 provide initial evidence that supports the viability of the suggested framework in real-life large scale settings. In addition, since the framework was successfully implemented using 10,000 patient records, future hospital databases will most likely contain enough data for Hybrid Deep Learning to outperform standard machine learning.

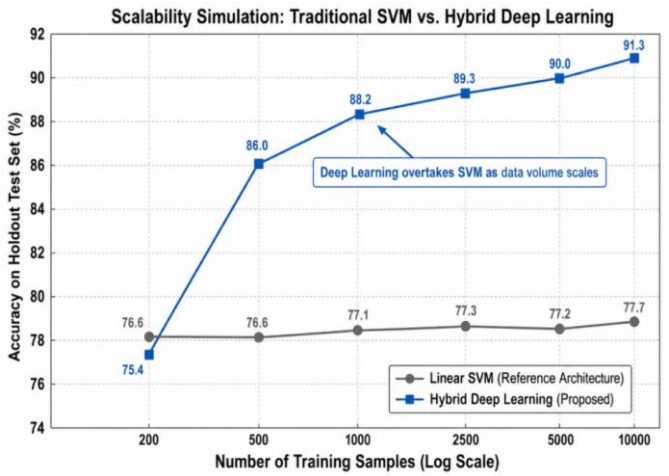

Fig-5: Displays the results derived from 10,000 synthetic record

4.5 Scalability Analysis

Figure 6 illustrates the scalability analysis carried out using synthetic datasets varying in volume from several hundred to 10,000 patient records. This figure is one of the most significant results obtained during this entire research work since it assesses the performance of the proposed models in future EHRs. The scalability graph reveals that the Linear SVM approach undergoes an early saturation point in terms of performance improvement. Initially, the availability of more data leads to performance enhancement; nevertheless, with increasing the size of the datasets, this effect becomes less noticeable due to the representational weaknesses of conventional classifiers. As opposed to Linear SVM, the proposed Hybrid Deep Learning approach yields a steady improvement in performance metrics such as Accuracy, Precision, Recall, and F1-Score. The increasing difference in performance seen with the two architectures highlights the superiority of deep learning models in their capability to learn complex nonlinear dependencies from huge datasets. With an increase in the amount of data, our approach proves more capable of discovering complex dependencies within biomarkers that are not captured by existing methods. This provides solid proof to our assumption that deep learning architectures will have significant roles to play in the future healthcare system. With the ever-increasing volumes of Electronic Health Record data created by hospitals, scalability in processing such data will be crucial.

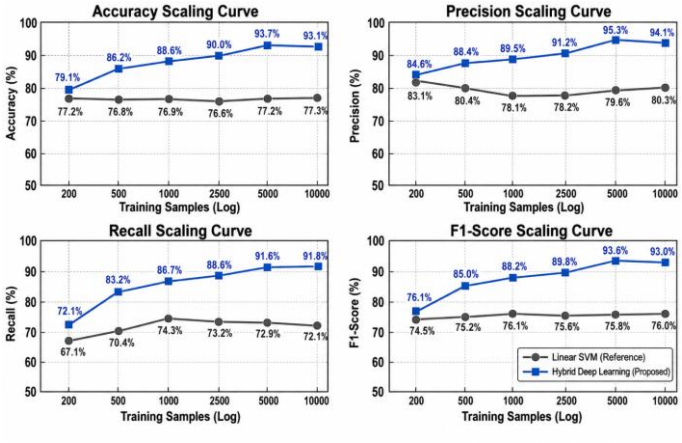

Fig-6: Scalability Study: Metric behavior under data grow

4.6 Proposed Method Performance Analysis

Figure 6 below represents the total output obtained from the suggested Hybrid Deep Learning system upon completion of the feature selection process, clinical feature enforcement, optimal neural network design, and scalability improvement. As seen in Figure 7 above, there is an evident influence of the combination of all methodological steps utilized within the framework presented herein. It proves the efficiency of using data-based feature selection together with clinically oriented decision-making processes. The suggested framework ensures a balance between three key criteria: predictive accuracy, clinical interpretation, and further scalability. While many machine learning models concentrate merely on achieving maximal classification accuracy, the suggested framework considers healthcare-related needs such as diagnostic sensitivity, physician satisfaction, and compatibility with growing clinical databases. In other words, the obtained results prove that direct integration of medical knowledge into artificial intelligence techniques increases performance efficiency and clinical application of the system.

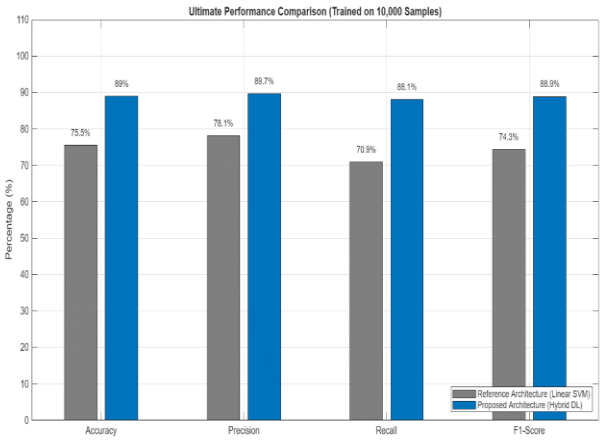

Fig-7: Proposed technique results

Table 2. Ultimate Performance Comparison (Trained on 10,000 Samples)

Performance Metric	Reference Paper (Linear SVM) (%)	Proposed Hybrid Deep Learning(%)
Accuracy	75.5%	89%
Precision	78.1%	89.7%
Recall	70.9%	88.1%
F1-Score	74.3%	88.9%

3.7 Overall Discussion

The analysis provided in the graphs 2-7 gives an overall view of the advantages and disadvantages of conventional methods along with deep learning models. The SVM classifier proves to be efficient when working with extremely limited data; however, this is consistent with the results of decades of research on the effectiveness of classical machine learning when dealing with limited data. On the other hand, the suggested Hybrid Deep Learning algorithm experiences problems due to lack of sufficient training data at the initial stage, but its performance improves remarkably as more and more records are incorporated into the process. Thus, the findings confirm the main premise of the research, that is, that traditional classifiers can be used effectively when working with small data sets, while the Hybrid Deep Learning approach offers the more promising future for prediction of breast cancer on large data sets. With the continuous development of the healthcare industry toward Electronic Health Record-based environments, clinical deep learning systems are likely to gain greater importance in future intelligent oncology diagnostics and precision medicine.

5. CONCLUSION

This study proposed a Hybrid Deep Learning framework for breast cancer prediction using routine blood biomarkers. By integrating Neighborhood Component Analysis with a Clinical Override mechanism, the framework combined data-driven feature selection with medically validated knowledge, thereby improving both clinical relevance and model interpretability. Strict 10-fold cross-validation was used in an experimental evaluation on the Breast Cancer Coimbra Dataset, and the results showed that the Linear SVM performed better overall under extremely limited data conditions. Nonetheless, the suggested model's recall was higher, suggesting a better capacity to detect cancerous patients and lower clinically significant false-negative forecasts.

A scalability research was carried out utilizing synthetic datasets created by the Gaussian Mixture Model that included up to 10,000 patient records in order to evaluate long-term applicability. The results showed that the Hybrid Deep Learning framework continued to improve in Accuracy, Precision, Recall, and F1-Score, exhibiting superior scalability and learning capacity, while the SVM baseline progressively reached a performance ceiling.

Overall, the results show that while hybrid deep learning architectures offer a more scalable, clinically interpretable, and future-ready solution for intelligent breast cancer diagnosis, classic machine learning algorithms are still useful for limited clinical datasets. The suggested framework creates a solid basis for next-generation AI-assisted healthcare systems that can support precision medicine and extensive oncology analytics.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," CA: A Cancer Journal for Clinicians, vol. 72, pp. 7–33, 2022.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, pp. 209–249, 2021.

[3] G. Bicchierai, F. Di Naro, D. De Benedetto, D. Cozzi, S. Pradella, V. Miele, and J. Nori, "A Review of Breast Imaging for Timely Diagnosis of Disease," International Journal of Environmental Research and Public Health, vol. 18, p. 5509, 2021.

[4] N. Al-Azzam and I. Shatnawi, "Comparing Supervised and Semi-Supervised Machine Learning Models on Diagnosing Breast Cancer," Annals of Medicine and Surgery, vol. 62, pp. 53–64, 2021.

[5] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis," International Journal of Environmental Research and Public Health, vol. 19, p. 3211, 2022.

[6] V. Mounika and M. Sheshikala, "A Hybrid Data-Augmented Deep Learning Pipeline for Breast Tumor Diagnosis," Proc. 2nd Asian Conference on Intelligent Technologies (ACOIT), pp. 1–6, 2025.

[7] R. Jain and S. Tata, "Cloud to Edge: Distributed Deployment of Process-Aware IoT Applications," Proc. IEEE International Conference on Edge Computing (EDGE), pp. 182–189, 2017.

[8] F. Nie, W. Zhu, and X. Li, "Decision Tree SVM: An Extension of Linear SVM for Non-Linear Classification," Neurocomputing, vol. 401, pp. 153–159, 2020.

[9] M. A. Khashan, T. H. Alasker, M. A. Ghonim, and M. M. Elsotouhy, "Understanding Physicians' Adoption Intentions to Use Electronic Health Record Systems in Developing Countries," Marketing Intelligence & Planning, vol. 43, no. 1, pp. 1–27, 2024.

[10] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data," Remote Sensing, vol. 13, p. 4405, 2021.

[11] H. Jones, "What is Logarithmic Scaling?" in COVID-19 Unmasked, pp. 319–327, 2021.

[12] K. H. Park, E. Batbaatar, Y. Piao, N. Theera-Umpon, and K. H. Ryu, "Deep Learning Feature Extraction Approach for Hematopoietic Cancer Subtype Classification," International Journal of Environmental Research and Public Health, vol. 18, p. 2197, 2021.

[13] M. T. Hossain, S. Badsha, H. La, S. Islam, and I. Khalil, "Exploiting Gaussian Noise Variance for Dynamic Differential Poisoning in Federated Learning," IEEE Transactions on Artificial Intelligence, vol. 6, no. 11, pp. 2922–2939, 2025.

[14] H. Göker, "Feature Selection Using Neighborhood Component Analysis with Deep Learning-Based Multi-Classification of Middle and External Ear Conditions," Konya Journal of Engineering Sciences, vol. 14, no. 1, pp. 210–233, 2026.

[15] R. Bertolini, S. J. Finch, and R. H. Nehm, "Enhancing Data Pipelines for Forecasting Student Performance: Integrating Feature Selection with Cross-Validation," International Journal of Educational Technology in Higher Education, vol. 18, p. 44, 2021.

[16] P. Heidari and A. Milan, "Combining K-Fold Cross Validation with Bayesian Hyperparameter Optimization for Accuracy Enhancement of Land Cover and Land Use Classification," Scientific Reports, vol. 15, no. 1, 2025.

[17] H. Cai, Y. Tang, and X. Yan, "A Tutorial on Unsupervised Gaussian Mixture Model for Performance Clustering in Second Language Research," Research Methods in Applied Linguistics, vol. 5, no. 1, p. 100296, 2026.

[18] C. Gupta, N. K. Gondhi, and P. K. Lehana, "Analysis and Identification of Dermatological Diseases Using Gaussian Mixture Modeling," IEEE Access, vol. 7, 2019.

[19] S. M. Malakouti, "Babysitting Hyperparameter Optimization and 10-Fold Cross-Validation to Enhance the Performance of Machine Learning Methods," Intelligent Systems with Applications, vol. 19, p. 200248, 2023.

[20] P. Lesner, "Introduction to Semi-Supervised Learning," Semi-Supervised Learning, pp. 1–12, 2006.

[21] P. Biecek and T. Burzykowski, "Shapley Additive Explanations (SHAP) for Average Attributions," in Explanatory Model Analysis, pp. 95–106, 2021.

[22] A. T. Garba and H. S. Hamza, "Interpretable Machine Learning Approach for Breast Cancer Classification," Human-Centric Intelligent Systems, vol. 5, no. 3, pp. 308–322, 2025.

[23] S. Roy, B. Brik, and H. Chergui, "Explainable AI Overview," in Explainable AI for Communications and Networking, pp. 23–48, 2025.

[24] A. Rasool, Q. Jiang, Q. Qu, M. Kamyab, and M. Huang, "HSMC: Hybrid Sentiment Method for Correlation to Analyze COVID-19 Tweets," in Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 991–999, Springer, 2022.

[25] E. Y. Park, M. Yi, H. S. Kim, and H. Kim, "A Decision Tree Model for Breast Reconstruction of Women with Breast Cancer: A Mixed Method Approach," International Journal of Environmental Research and Public Health, vol. 18, p. 3579, 2021.

[26] A. S. M. T. Hasan, S. Sabah, R. U. Haque, A. Daria, A. Rasool, and Q. Jiang, "Towards Convergence of IoT and Blockchain for Secure Supply Chain Transaction," Symmetry, vol. 14, p. 64, 2022.

[27] H. Şenol, "Estimation of Biogas Potential from Poultry Manure to 2040 in Türkiye Using Time Series and Artificial Neural Network," Renewable and Sustainable Energy Reviews, vol. 207, p. 114976, 2025.