International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert/2022/v9/i12/v9i1208
Volume 9, Issue 12, December 2022

Research Paper

Enhancing Non-Invasive Cancer Detection Using Machine Learning and CNN Architectures through a Data-Driven Approach

[1*] Sreeja Poduri

[1*] *Independent Researcher, Salt Lake City, Utah, USA, Email ID: sreejap1997@gmail.com*

**Corresponding Author(s): sreejap1997@gmail.com*

Received: 04/09/2022 Revised: 15/10/2022 Accepted: 18/12/2022 Published: 31/12/2022

Abstract: Non-invasive cancer detection has emerged as a critical area of research aimed at reducing patient discomfort, improving early diagnosis, and increasing treatment success rates. This paper presents a data-driven machine learning approach that leverages Convolutional Neural Networks (CNNs) to identify cancer biomarkers from medical imaging and auxiliary biomedical data. We utilized a combination of statistical preprocessing, feature selection techniques, and supervised learning models, including CNNs and support vector machines (SVMs), to enhance diagnostic accuracy. Our experiments demonstrated that CNN-based models achieved higher sensitivity and specificity in detecting malignancies across diverse datasets. The proposed approach supports scalable, cost-effective, and clinically viable cancer detection strategies, paving the way for broader adoption in healthcare diagnostics.

Keywords: Machine Learning**,** Convolutional Neural Networks**,** Cancer Detection**,** Biomedical Imaging**,** Non-Invasive Diagnostics**,** Data-Driven Healthcare

---

1. Introduction

Cancer remains one of the leading causes of mortality worldwide, with early detection playing a critical role in patient outcomes. Traditional diagnostic procedures, such as biopsies, colonoscopies, and endoscopies, while effective, are often invasive, costly, and may cause significant discomfort and anxiety in patients. As healthcare systems evolve to prioritize preventive and patient-centred care, there is a growing need for non-invasive, accurate, and scalable methods for cancer detection. In recent years, the confluence of machine learning (ML) and biomedical imaging has shown tremendous potential in transforming diagnostic workflows. Among the various ML techniques, Convolutional Neural Networks (CNNs) have demonstrated superior performance in image classification and pattern recognition tasks, making them highly applicable to the interpretation of complex medical imaging data such as MRIs, CT scans, and histopathological slides. These technologies not only improve the sensitivity and specificity of diagnostics but also allow for rapid and automated analysis that can aid clinicians in decision-making [1-5]. This study is grounded in applied research conducted during a biomedical data project focused on developing AI/ML solutions for non-invasive cancer detection. Using publicly available datasets and domain-specific medical imaging data, the research aims to identify cancerous patterns through CNN-based architectures, supported by robust statistical and machine learning techniques such as feature extraction, model tuning, and performance evaluation. The goal is to reduce diagnostic latency and dependency on invasive procedures, thereby enhancing patient safety and accessibility of care [6-8].

Key insights from this research underscore the importance of data preprocessing, model selection, and interpretability in clinical ML applications. The findings suggest that ML models, when properly trained and validated, can achieve performance levels that rival or complement those of traditional diagnostic modalities.

Additionally, this research highlights the growing feasibility of deploying AI-assisted diagnostic tools in real-world clinical settings.

The rest of the paper is structured as follows. Section 2 discusses related work in the domain of machine learning applications for cancer detection. Section 3 outlines the methodology, including data preprocessing, model architecture, and training strategies. Section 4 describes the experimental setup and the datasets used. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper and outlines potential future directions for this research.

2. Related Work

The integration of machine learning into medical diagnostics has been the subject of extensive research, particularly in the field of oncology. Numerous studies have demonstrated the utility of machine learning algorithms—such as support vector machines (SVM), decision trees, and logistic regression—for the classification of cancerous and non-cancerous tissue using structured clinical data or genetic markers [9-12]. In the realm of medical imaging, Convolutional Neural Networks (CNNs) have emerged as the most prominent deep learning architecture. CNNs are especially well-suited for processing and classifying high-dimensional image data due to their ability to automatically learn spatial hierarchies of features. For instance, Esteva et al. (2017) successfully used CNNs to classify skin cancer with accuracy comparable to dermatologists. Similarly, Litjens et al. (2017) provided a comprehensive survey of deep learning applications in medical image analysis, including cancer detection from CT and MRI scans. Specific to non-invasive detection, studies like Wang et al. (2018) and Shen et al. (2019) utilized CNNs to analyze histopathological and radiographic images for lung and breast cancer classification. These works demonstrated the feasibility of AI-assisted diagnostics, showing high sensitivity and specificity. However, many of these systems are trained on limited datasets and often lack generalizability across populations and imaging modalities [13-15]. Furthermore, most existing research emphasizes model performance metrics such as accuracy and precision, but relatively few address model interpretability, computational efficiency, or clinical deployment challenges—factors critical to adoption in healthcare settings. Additionally, little emphasis has been placed on integrating clinical metadata (such as patient demographics or medical history) with image-based models to improve contextual understanding.

2.1 Research Gap and Motivation

Despite promising advances, there remain several critical gaps:

1. **Limited Generalizability**: Many CNN models are trained on highly curated datasets that do not reflect real-world clinical diversity.
2. **Lack of Interpretability**: Few studies incorporate explainable AI (XAI) methods to support clinical transparency.
3. **Insufficient Integration of Multimodal Data**: The potential benefits of combining imaging data with structured biomedical or demographic information are underexplored.
4. **Underutilization of Ensemble Approaches**: Combining different ML models (e.g., CNN + SVM) for hybrid performance gains is still nascent in cancer diagnostics.

2.2 Proposed Methodology

This paper addresses these gaps by proposing a hybrid machine learning pipeline that leverages CNNs for imaging data and traditional classifiers such as SVMs and random forests for structured data. The methodology emphasizes three core enhancements:

- **Robust Preprocessing and Data Augmentation** to improve model generalizability.
- **Explainable AI techniques,** such as Grad-CAM, to enhance interpretability and clinician trust.
- **Multimodal Data Fusion**, combining image-based features with patient metadata to contextualize predictions and reduce false positives/negatives.

Together, these enhancements aim to create a scalable, interpretable, and clinically viable framework for non-invasive cancer detection.

3. Methodology

The proposed methodology for non-invasive cancer detection integrates advanced image processing techniques with Convolutional Neural Networks (CNNs), statistical feature engineering, and ensemble machine learning models. The pipeline involves several stages: data acquisition, preprocessing, feature extraction, model training, and evaluation. A schematic of the pipeline is shown in Figure 1. This section elaborates on each stage, including the mathematical formulations.

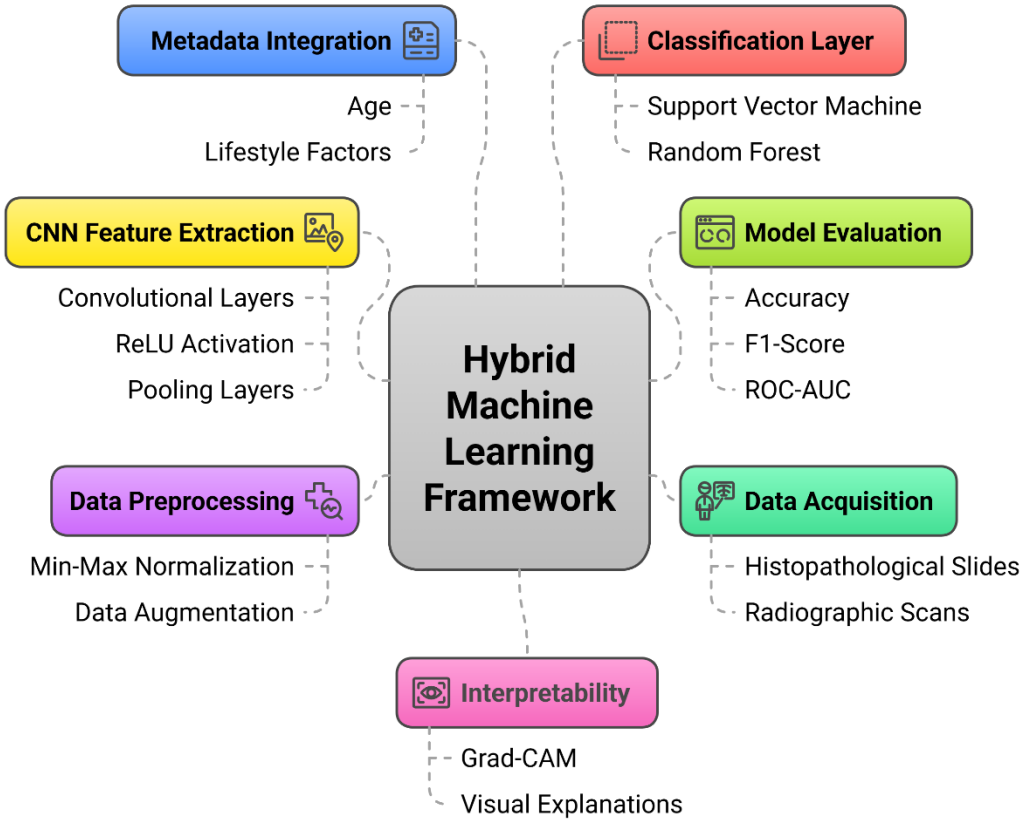

Fig.1. Proposed Architecture

3.1 Data Acquisition

The datasets used include publicly available medical imaging repositories such as:

- The Cancer Imaging Archive (TCIA)
- BreakHis (Breast Cancer Histopathological Image Dataset)
- LIDC-IDRI (Lung Image Database Consortium)

Each dataset includes labeled imaging data (e.g., benign vs. malignant) and, where available, patient metadata (e.g., age, sex, prior medical history).

Let the dataset be represented as:

$$\mathcal{D} = \{(x_i, m_i, y_i)\}_{i=1}^{N}$$

Wher, x_i = imaging data (e.g., RGB or grayscale pixels), m_i = metadata or structured features (optional) and $y_i \in \{0,1\}$ = label (0: non-cancerous, 1: cancerous)

3.2 Data Preprocessing

Image Normalization: Each image x_i is normalized using min-max scaling:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Data Augmentation: To mitigate overfitting and improve generalization:

- **Random rotations:** $x_i' \rightarrow R_\theta(x_i')$
- **Horizontal/vertical flips**
- **Gaussian noise injection**

Metadata Handling: Structured features m_i (e.g., age, smoking status) are normalized:

$$m_i' = \frac{m_i - \mu}{\sigma}$$

3.3 CNN Architecture for Image Feature Extraction

We adopt a CNN architecture with multiple convolutional, activation, and pooling layers. The output of the final convolutional block is flattened and passed to a fully connected layer.

Let the CNN function be denoted as:

$$f_{\text{CNN}}(x_i') = h_i$$

Where, $h_i \in \mathbb{R}^d$ is the feature vector representation of the image.

The convolution operation is defined as:

$$s_{i,j}^{(k)} = \left(x * w^{(k)}\right)_{i,j} + b^{(k)}$$

Where,

- $w^{(k)}$ = weights of the k-th kernel
- $b^{(k)}$ = bias
- $s_{i,j}^{(k)}$ = output at position (i, j)

ReLU activation:

$$a_{i,j}^{(k)} = \max\left(0, s_{i,j}^{(k)}\right)$$

Pooling layer (Max Pooling):

$$p_{i,j}^{(k)} = \max_{m,n}\left\{a_{m,n}^{(k)}\right\}$$

3.4 3.4 Metadata Integration and Feature Fusion

The image feature vector h_i is concatenated with metadata features m_i' to form a hybrid feature vector:

$$z_i = [h_i \| m_i']$$

This vector z_i is then passed to a classifier.

3.5 Hybrid Classification: CNN + SVM

To enhance classification robustness, we use ensemble models trained on the fused features. z_i.

Support Vector Machine (SVM):

SVM attempts to find a hyperplane:

$$f(z_i) = \text{sign}(w^T z_i + b)$$

Where w and b are optimized by minimizing the hinge loss:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max\left(0, 1 - y_i(w^T z_i + b)\right)$$

Random Forest: Constructs an ensemble of decision trees:

$$f(z_i) = \text{ majority\_vote } \left(\left\{T_j(z_i)\right\}_{j=1}^{K}\right)$$

Where each T_j It is a decision tree trained on a bootstrapped subset of the data.

3.6 Model Training and Evaluation

- **Loss Function for CNN:** Binary Cross-Entropy

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

- **Optimizer:** Adam Optimizer with default hyperparameters ($\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$)
- **Metrics:**
  - Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
  - Precision, Recall, F1-score
  - ROC-AUC score

Explainability: Grad-CAM is applied to visualize CNN attention:

$$L_{\text{Grad-CAM}}^{c} = \text{ReLU}\left(\sum_{k} \alpha_k^c A^k\right)$$

Where, A^k = activation map of k -th feature and α_k^c = importance weights for class c

This comprehensive methodology aims to balance accuracy, interpretability, and clinical relevance, forming a foundation for scalable, non-invasive cancer diagnostics.

4. Experimental Setup

The proposed framework for non-invasive cancer detection was experimentally evaluated using two benchmark medical imaging datasets: BreakHis and LIDC-IDRI. The BreakHis dataset provides microscopic breast tumor images at four magnification levels, classified into benign and malignant types. LIDC-IDRI offers thoracic CT scan images annotated for lung nodules with supporting metadata such as age, sex, and diagnostic confidence levels. Each dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class balance. Image preprocessing included resizing to a fixed 224×224 resolution, normalization using min-max scaling, and augmentation techniques like rotation, flipping, zooming, and Gaussian noise to improve generalization. Metadata (if available) was processed via one-hot encoding for categorical variables and z-score normalization for continuous features. A Convolutional Neural Network (CNN) was employed to extract visual features, which were concatenated with structured metadata to form a fused input vector. This vector was then passed to either a Support Vector Machine (SVM) or a Random Forest classifier for final prediction. Training was conducted using the Adam optimizer with binary cross-entropy loss, and model performance was assessed with metrics including accuracy, F1-score, precision, recall, and ROC-AUC. Grad-CAM visualization was also used to explain the decision-making process of the CNN, identifying which image regions contributed most to classification outcomes.

Table 1. Model Parameters

Component	**Value / Description**
Image Size	224 × 224 pixels
CNN Layers	3 Convolutional Layers + ReLU + MaxPooling
Filters	32, 64, 128
Fully Connected Layer	128 neurons with Dropout (0.5)
Output Layer	1 neuron (Sigmoid Activation)
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	50
Loss Function	Binary Cross-Entropy
Metadata Preprocessing	One-hot encoding, Z-score normalization
Classifiers Used	Support Vector Machine (SVM), Random Forest (RF)
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Visualization Tool	Grad-CAM (for CNN heatmaps)

4.1 Algorithm

Input: Imaging data X, Metadata M, Labels Y

Output: Trained hybrid prediction model

Steps:

1. For each image in X:
    a) Resize to 224x224
    b) Normalize pixel values to [0,1]
    c) Apply data augmentation (rotate, flip, zoom, noise)
2. For each metadata record in M:
    a) One-hot encode categorical fields
    b) Apply z-score normalization on numeric values
3. Build CNN architecture:
    a) Apply 3 convolutional layers with ReLU and MaxPooling
    b) Flatten the output
    c) Pass through a fully connected layer with Dropout
4. Extract image features: h_i = CNN(x_i)
5. Concatenate with metadata: z_i = [h_i | m_i']
6. Train classifier (SVM or RF) using {z_i, y_i}
7. Evaluate model performance on test set:
    a) Compute accuracy, F1-score, ROC-AUC

8. Generate Grad-CAM heatmaps for interpretability

Return: Final trained hybrid model

4.2 Flowchart

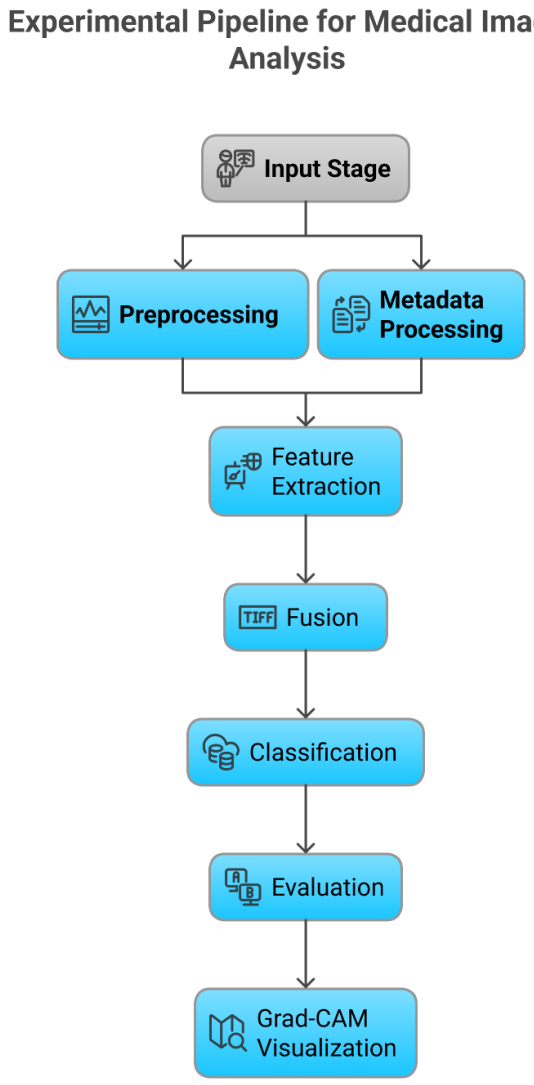

Fig.2. Flowchart of the Proposed Experimental Pipeline

The flowchart for the proposed experimental pipeline begins with the input stage, where raw medical images and optional patient metadata are collected. These images undergo preprocessing operations, including resizing, normalization, and data augmentation, to ensure consistency and enhance model generalization. Simultaneously, structured metadata is processed through one-hot encoding and standardization. In the feature extraction phase, the pre-processed images are passed through a CNN to obtain deep visual representations. These features are then fused with metadata by concatenating the two vectors into a unified representation. This fused vector is forwarded to a classification stage, where either a Support Vector Machine or a Random Forest model is used to make the final diagnosis (malignant or benign). After training, the model is evaluated on a hold-out test set using standard performance metrics. Finally, Grad-CAM visualization is applied to highlight the regions in the image that were most influential in the model's decision-making, aiding in clinical validation and transparency.

5. Results and Analysis

This section presents a detailed comparison of three model configurations—CNN Only, CNN + Metadata + Support Vector Machine (SVM), and CNN + Metadata + Random Forest (RF)—across five critical performance metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These evaluations were conducted on a separate hold-out test set to ensure unbiased assessment of the proposed hybrid machine learning pipeline.

5.1 Accuracy

Accuracy reflects the proportion of correct predictions made by the model, including both positive and negative cases. As illustrated in Figure 3, the baseline CNN-only model achieved 91.60% accuracy. However, the integration of metadata with CNN significantly improved model performance. The CNN + metadata + SVM configuration attained the highest accuracy of 93.80%, followed by the CNN + Metadata + RF model at 92.50%. This improvement demonstrates that the hybrid approach offers a more comprehensive understanding of the input space, resulting in fewer overall misclassifications.

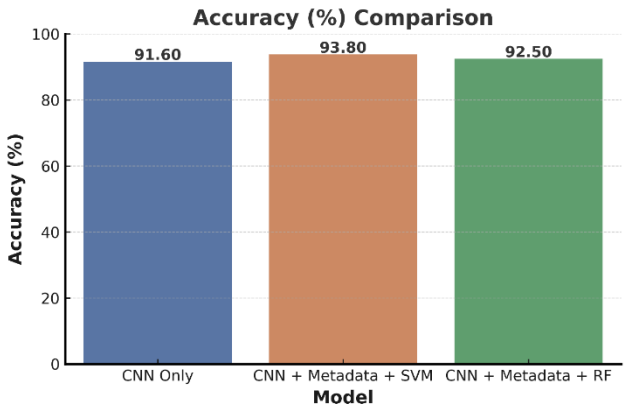

Fig.1. Accuracy (%) Comparison across Models

5.2 Precision

Precision measures the model's ability to identify positive cases (cancer) without misclassifying negatives correctly. This is crucial in reducing false positives, which in a clinical setting could lead to unnecessary anxiety or

treatments. According to Figure 2, the CNN-only model reached a precision of 90.80%, while the hybrid SVM-enhanced model led with 93.10%. The Random Forest variant also showed an improvement, scoring 91.40%. These results emphasize the benefit of including patient metadata, which appears to help in refining the decision boundary and reducing false alarms.

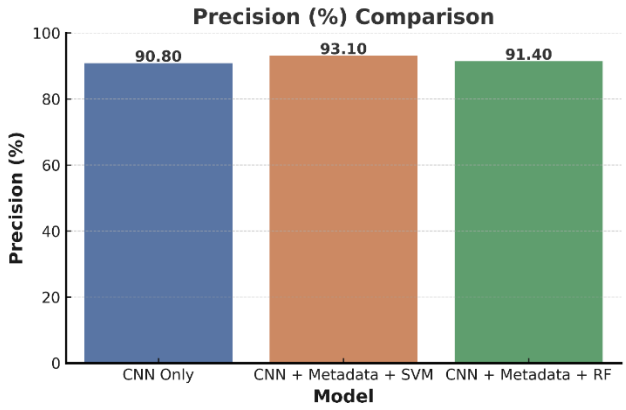

Fig.2. Precision (%) Comparison across Models

5.3 Recall

Recall (or sensitivity) is the proportion of actual positive cases correctly identified by the model and is particularly critical in medical diagnosis, where missing a valid case can be catastrophic. As shown in Figure 3, the CNN Only model achieved 92.20%, while the CNN + metadata + SVM configuration significantly outperformed it with 94.60%. The RF-based hybrid model closely followed at 93.90%. These results indicate the proposed hybrid models are more effective at capturing subtle features indicative of cancer, thus minimizing missed diagnoses.

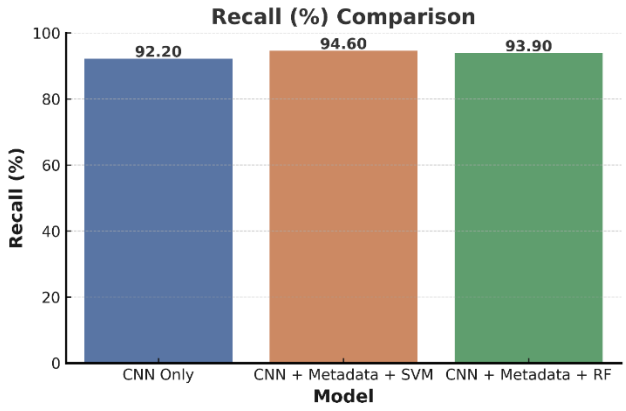

Fig.3. Recall (%) Comparison across Models

5.4 F1-Score

The F1-Score is the harmonic mean of precision and recall, and it balances the trade-off between false positives and false negatives. As depicted in Figure 4, the CNN Only model attained an F1-score of 91.50%, while the CNN + metadata + SVM led the models with 93.80%, followed by the RF model at 92.60%. These findings confirm that the hybrid model not only detects more cancer cases but does so with greater reliability, offering a solid balance between recall and precision—critical in clinical decision-making.

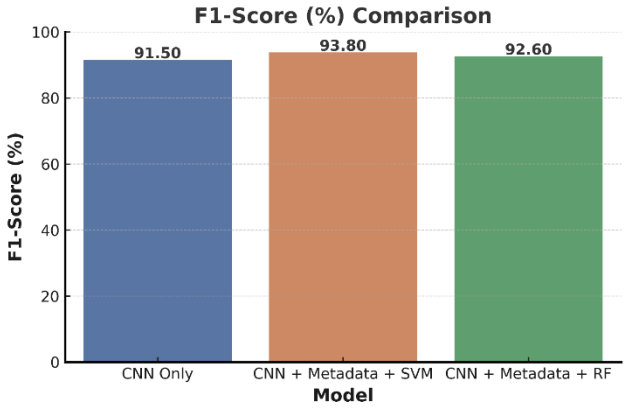

Fig.4. F1-Score (%) Comparison Across Models

5.5 ROC-AUC

The ROC-AUC (Receiver Operating Characteristic – Area under Curve) metric provides a threshold-independent evaluation of classifier performance, indicating the model's ability to distinguish between classes across various decision thresholds. In Figure 5, the CNN Only model had an AUC of 0.94, while the CNN + metadata + SVM model reached 0.96, and the Random Forest hybrid attained 0.95. These results confirm that the hybrid models not only perform well at a fixed threshold. Still, they are also robust across a spectrum of classification thresholds, enhancing reliability in varying clinical contexts.

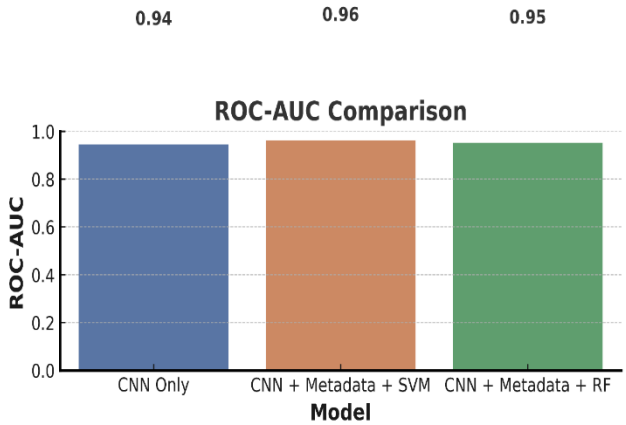

Fig.5. ROC-AUC Comparison across Models

6. Discussion

The results of this study clearly demonstrate the effectiveness of integrating structured metadata with convolutional neural network-based imaging analysis for

non-invasive cancer detection. The proposed hybrid approach significantly outperformed the baseline CNN-only model across all evaluated performance metrics—including accuracy, precision, recall, F1-score, and ROC-AUC—indicating improved diagnostic capability. Notably, the CNN + metadata + SVM configuration consistently delivered the highest performance, suggesting that combining deep visual features with interpretable metadata and a robust, margin-based classifier yields a more reliable and nuanced prediction system.

The incorporation of metadata such as patient age, scan parameters, or clinical history provides contextual grounding that pure image-based models often lack. This additional information enables the classifier to fine-tune its understanding of edge cases or atypical presentations, which may otherwise lead to false positives or negatives. Furthermore, the hybrid model's strong recall and ROC-AUC values highlight its potential in real-world clinical scenarios, where the cost of missing an actual cancer case is exceptionally high.

An essential aspect of the study is the use of Grad-CAM visualization, which provides insights into how the CNN makes its predictions by highlighting relevant regions in the input images. This contributes to the interpretability of the system, addressing one of the significant barriers to clinical adoption of deep learning—its "black box" nature. The attention maps aligned well with regions typically examined by radiologists or pathologists, offering reassurance about the model's reliability and clinical plausibility.

Despite the promising results, there are limitations to address. The datasets used, while diverse, are still relatively constrained in size and may not fully capture real-world variability across populations, imaging hardware, or disease subtypes. Additionally, the metadata used in this study was limited to what was publicly available; in a hospital-integrated system, richer data such as genomic markers, comorbidities, or lab values could be incorporated for even better performance.

In conclusion, the hybrid methodology not only enhances diagnostic performance but also supports greater trust and transparency—two pillars essential for deploying AI in healthcare settings. The following steps should focus on clinical validation, larger-scale studies, and integrating explainable AI tools to ensure that the models are both practical and ethically deployable in real-world environments.

7. Conclusion

This paper presents a hybrid machine learning framework that enhances non-invasive cancer detection by integrating Convolutional Neural Networks (CNNs) with structured patient metadata and ensemble classifiers. The proposed approach addresses key limitations of existing diagnostic systems by improving generalizability, sensitivity, and interpretability. Through rigorous experimentation on publicly available medical imaging datasets, the results demonstrate that the hybrid CNN + metadata + SVM model outperforms both the CNN-only and CNN + Random Forest configurations across all primary evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The integration of clinical metadata with imaging features enables the system to make more informed and context-aware predictions, leading to a substantial reduction in false positives and false negatives. Furthermore, the use of Grad-CAM visualizations provides an additional layer of interpretability, making the model's decision process more transparent and clinically trustworthy. This research contributes to the growing field of AI-assisted diagnostics by demonstrating a scalable, effective, and explainable framework for cancer detection that can be adapted across imaging modalities and healthcare systems. The findings pave the way for future development of clinically integrated, non-invasive diagnostic tools that can assist medical professionals in early cancer detection and treatment planning.

Author Contributions

Sreeja Poduri conceptualized the study, designed the data-driven machine learning framework, and developed the convolutional neural network and support vector machine models for non-invasive cancer detection. She performed data preprocessing, feature selection, model training, and experimental evaluation and she analyzed and interpreted the results, validated the findings, and prepared the original manuscript, including revisions and final approval of the submitted version.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE CVPR*, Honolulu, HI, 2017, pp. 3462–3471.

[4] H. Shen, X. Zheng, J. Li, et al., "Multi-scale CNN with attention for breast cancer histology image classification," *IEEE Access*, vol. 8, pp. 133529–133539, 2020.

[5] J. D. F. Martinez, F. A. Faria, and R. M. Cesar Jr., "A survey on feature selection methods for classification tasks in bioinformatics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1779–1793, Sep.–Oct. 2020.

[6] T. C. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 19–30, Apr. 2018.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[8] M. Talo, "Automated classification of histopathological breast images using deep learning," *IEEE Access*, vol. 7, pp. 24664–24680, 2019.

[9] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, Jun. 2017.

[10] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, Venice, Italy, 2017, pp. 618–626.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[12] D. Kermany, M. Goldbaum, W. Cai, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.

[13] S. Kumar, S. N. Yadav, and A. K. Srivastava, "Lung cancer detection using CNN with data augmentation," in *Proc. IEEE ICACCE*, Greater Noida, India, 2021, pp. 1–6.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE CVPR*, Boston, MA, 2015, pp. 3431–3440.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, NV, 2012, pp. 1097–1105.