Ensemble Learning vs. Deep Learning for Breast Cancer Classification from Biopsy Data: A Comparative Study with Explainable AI Interpretability

Dr. P. Bhargavi Devi [1], V. Prema (Corresponding Author), K. Pandi Meena[3], B.Chithra[4]
[1]Assistant Professor, SRM Institute of Science and Technology Ramapuram, bhargavp@srmist.edu.in
[2]Assistant Professor, Alpha Arts and Science College, premaveluchamy78@gmail.com
[3]Assistant Professor, SRM Institute of Science and Technology Ramapuram, pandimek@srmist.edu.in
[4]Assistant Professor, SRM Institute of Science and Technology Ramapuram, chitrab@srmist.edu.in

To Cite this Article

Dr. P. Bhargavi Devi, V. Prema, K. Pandi Meena, B. Chithra," **Ensemble Learning vs. Deep Learning for Breast Cancer Classification from Biopsy Data: A Comparative Study with Explainable AI Interpretability"** ***Musik In Bayern, Vol. 89, Issue 11, Nov 2024, pp17-30***

Article Info

Received: 24-10-2024 Revised: 06-11-2024 Accepted: 14-11-2024 Published: 24-11-2024

Abstract:

In this research, we investigate the efficacy of Ensemble Learning and Deep Learning methodologies for the classification of breast cancer using biopsy data. Ensemble techniques, including Random Forest, Gradient Boosting, and Stacking, are employed in the Machine Learning (ML) domain, while a Feed Forward Neural Network is utilized in the Deep Learning (DL) domain. The study focuses on benign/malicious classification, crucial for accurate diagnosis and treatment planning. Through rigorous hyperparameter tuning and performance evaluation using the log loss metric, we observe that ML models, particularly Gradient Boosting Classifier and Random Forest, exhibit robust performance in both accuracy and log loss metrics. In contrast, DL models initially demonstrate competitive performance but face challenges such as overfitting, as evidenced by a slight decrease in performance after early stopping. Our findings suggest that ML models, with their ensemble techniques and simpler architectures, present a more effective and robust approach for breast cancer classification from biopsy data. This research contributes to the growing field of computational biology and underscores the importance of selecting appropriate machine learning methodologies for biomedical applications in artificial intelligence (AI) research.

Keywords: Machine Learning, Deep Learning, Ensembling

Introduction

Breast cancer remains a significant public health concern worldwide, with substantial implications for patient outcomes and healthcare systems. Early and accurate classification of breast lesions as benign or malignant is paramount for timely diagnosis and effective treatment planning. In recent years, the intersection of computational biology and artificial intelligence (AI) has propelled advancements in breast cancer classification, with Machine Learning (ML) and Deep Learning (DL) emerging as powerful tools in this domain. Machine Learning techniques, particularly ensemble methods such as Random Forest, Gradient Boosting, and Stacking, have garnered attention for their ability to harness the collective intelligence of diverse base models. These ensemble techniques offer robustness and flexibility in handling complex datasets, making them well-suited for biomedical applications such as breast cancer classification. On the other hand, Deep Learning models, characterized by their multi-layered neural architectures, have demonstrated remarkable capabilities in extracting intricate patterns and features from raw data. Feed Forward Neural Networks, a fundamental DL architecture, have shown promise in various medical imaging tasks, including breast cancer diagnosis. Despite the growing interest in both ML and DL approaches for breast cancer classification, there remains a need to systematically compare their performance and identify the most effective methodologies for clinical application. This research aims to address this gap by conducting a comparative study of ML and DL models for breast cancer classification using biopsy data. Through rigorous experimentation and evaluation, we seek to elucidate the strengths and limitations of ensemble techniques and deep learning architectures in this critical domain.

The Paper organized as follows:

Section 3 explores the Related Work, Section 4 describes our Proposed Methodology, Section 5 outlines the Experimental Setup, Section 6 discusses the Techniques and Methodologies employed, Section 7 presents the Experimental Results, and Section 8 concludes the paper and discusses future work.

Related Study

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350

Khalid et al [1] compares several machine learning algorithms, including ensemble methods, for breast cancer detection using mammograms. they suggest that ensemble methods offer high accuracy and potential for clinical applications. Ali et al [2] explores the use of meta-learning to improve the performance of ensemble methods for breast cancer classification, they demonstrate that combining meta-learning and ensemble methods can significantly enhance accuracy. Bazazeh et al [3] compares various machine learning algorithms, including ensemble methods, for breast cancer diagnosis using biopsy data. It concludes that ensemble methods like Random Forest and Gradient Boosting achieve competitive performance compared to other techniques. Zakareya et al [4] proposes a novel deep learning model for breast cancer diagnosis using ultrasound images. It shows missing results in differentiating benign and malignant masses. Nasser et al [5] provides a comprehensive overview of deep learning methods used for breast cancer diagnosis, including convolutional neural networks (CNNs). It highlights the potential and challenges of using deep learning in this domain.

Several techniques for constructing heterogeneous ensembles are applied and comparatively evaluated by Kazmaier J, Van Vuuren JH [10] (2022) across four different domains of benchmark sentiment classification datasets, revealing median performance improvements over individual models. Alsayat A [11] (2022) established a sentiment analysis framework using deep learning and ensemble techniques tailored for COVID-19-related social media data. He used two key model development stages include creating a baseline classifier, such as an LSTM network, and proposing an ensemble model that combines various classifiers for enhanced performance Mohammed A, Kora R.[12] (2022) capitalizes on the variability of Tier-0 classifiers and their predictions, utilizing them to construct effective ensemble models in Tier-1 through the training of shallow meta-classifiers.

Phan et al [13] (2020) proposed a new approach based on a feature ensemble model related to tweets containing fuzzy sentiment by taking into account elements such as lexical, word-type, semantic, position, and sentiment polarity of words. and method has been experimented on with real data, and the result proves effective in improving the performance of tweet sentiment analysis in terms of the F1 score. word embeddings have been utilized as an alternative to the manual techniques [14][15] Fouad et al (2018) combined Bag of words with Lexicon, Emoticons and Part of speech (PoS) and gave better results with ensemble classifier.[16] The Continuous Bag-of-

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*350*
Words (CBOW) and Skip-gram models are two popular variants of the Word2Vec algorithm.[17] Pennington J (2014) introduced Glove and the extraction method outperforms CBoW and Skip Gram with 93.2 F1 score [18]

Methodology

The dataset comprising biopsy data would be preprocessed to ensure data quality and feature compatibility across models. Then, three ML techniques, namely Random Forest, Gradient Boosting Classifier, and Stacking, would be implemented with appropriate hyperparameter tuning. Additionally, a DL approach using a Feed Forward Neural Network architecture would be deployed. The models would undergo rigorous evaluation using performance metrics such as accuracy and log loss. The comparison between ML and DL models would involve analyzing their performance in terms of classification accuracy and predictive uncertainty. Finally, the impact of early stopping on the DL model's performance would be assessed to understand its role in preventing overfitting. This methodology ensures a systematic investigation of both ML and DL approaches, facilitating a comprehensive comparison of their effectiveness in breast cancer classification.

Experimental set up

The experiments were conducted on virtual machine in Google Colab platform The google platform provided a virtual machine with CPU resources, well-suited for running resource-efficient tasks. The 12 GB of RAM facilitated handling larger datasets and models. This platform provided cloud-based computing resources without the need for local hardware The programming environment employed for the experiments was Python. The dataset underwent a preprocessing phase to ensure its suitability for training and evaluating classification models. During this preprocessing, several key steps were applied to clean and prepare the data

Table 1: Dataset description

Nature of dataset	No of features	Class	No of observations
Breast cancer from Kaggle	30	1. M-Malignant 2. B-Benign	590

Preprocessing

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350

The dataset underwent a preprocessing phase to ensure its suitability for training and evaluating sentiment analysis models. During this preprocessing, several key steps were applied to clean and prepare the data. This included text normalization to handle variations in letter casing and punctuation, the removal of special characters and numerical values, and the elimination of stop words to reduce noise in the text. Additionally, techniques such as tokenization were employed to break down the text into individual words or tokens, enabling further analysis. Furthermore, any duplicate or irrelevant entries were removed to maintain data integrity. This cleaned and processed dataset, devoid of noisy or redundant information, was then utilized as the input for training and evaluating the sentiment analysis models, ensuring that the models could focus on the meaningful content of the reviews while minimizing the impact of irrelevant factors.

Ensembling in Machine Learning

Ensembling is a machine learning technique that combines the predictions of multiple individual models to improve overall performance. The idea is to leverage the diversity of different models to mitigate weaknesses and enhance predictive accuracy. Common ensembling methods include bagging, boosting, and stacking. In the context of breast cancer classification, ensembling techniques such as Random Forest, Gradient Boosting, and Stacking can be applied to combine the predictions of various machine learning models trained on biopsy data, thereby improving the accuracy and robustness of the classification system.

Feed Forward Neural Network (FNN) for Breast Cancer Classification:

A Feed Forward Neural Network (FNN) is a type of artificial neural network where connections between nodes do not form cycles. It consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of neurons, and information flows forward from the input layer through the hidden layers to the output layer without any feedback loops. FNNs are trained using gradient descent-based optimization algorithms to learn the complex patterns and relationships present in the data. In the context of breast cancer classification, an FNN can be trained on biopsy data to automatically extract relevant features and classify cases as benign or malignant based on learned patterns. However, careful tuning of hyperparameters and regularization techniques is necessary to prevent overfitting and ensure optimal performance. Additionally, early stopping techniques may be employed to halt training when performance on a

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*350*
validation set begins to degrade, thereby preventing overfitting and improving generalization ability. Overall, FNNs offer a powerful tool for breast cancer classification, leveraging their ability to capture intricate patterns in complex data sets.

Results and Discussions

Machine Learning Models:

The ML models, including Random Forest, Gradient Boosting Classifier, and Stacking, showcased exceptional performance in accurately classifying breast cancer cases from biopsy data. Random Forest achieved perfect accuracy (1.0) and a commendable log loss of -0.0201. Gradient Boosting Classifier excelled with an accuracy of 0.99 and the lowest log loss (-0.000459) among ML models. Stacking, while slightly lower in log loss, demonstrated strong classification capabilities with an accuracy of 0.98 and a log loss of -0.0144. Overall, these ML models exhibit promising potential for precise breast cancer diagnosis and treatment planning.

Deep Learning Model (Feed Forward Neural Network):

The feed forward neural network model demonstrated effective classification performance before early stopping, achieving an accuracy of 0.9666 and a relatively low log loss of 0.07909. However, after early stopping, there was a decline in performance, with accuracy decreasing to 0.9333 and log loss increasing to 0.1833. This indicates that while early stopping may have prevented overfitting, it also resulted in a slight decrease in predictive performance. Overall, the neural network exhibited good accuracy and low log loss initially, highlighting its potential for breast cancer classification, albeit with considerations for early stopping's impact on performance.

Table 2: Performance of Deep Learning Model

Model	**Log loss**	**Accuracy**
Feed forward Neural network (Before early stopping)	0.07909	0.9666
Feed forward Neural network (After early stopping)	0.1833	0.9333

Early stopping is a regularization technique commonly used in training deep learning models to prevent overfitting and improve generalization performance. It involves monitoring the model's

Musik in bayern
ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*350*
performance on a validation dataset during the training process and halting the training process when the performance on the validation set begins to degrade.

Fig1 :Loss Vs Epoch (Before Early stopping)

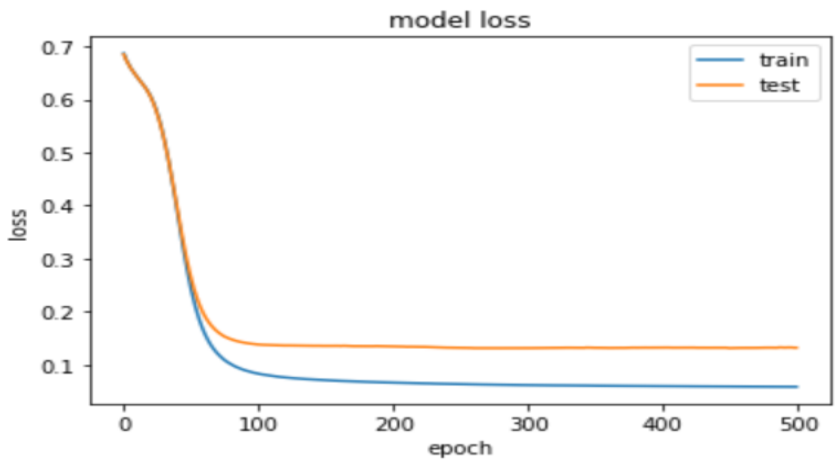
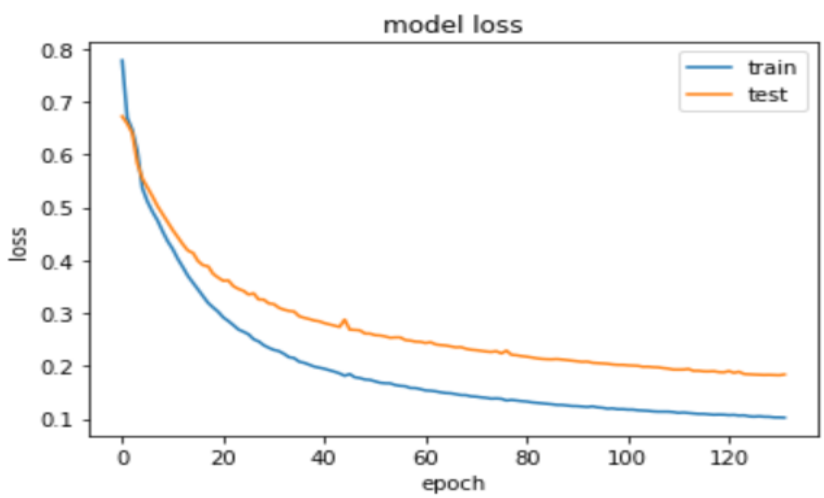


Fig 2: Loss Vs Epoch (Aftere Early stopping)



Comparative Interpretation:

ML models, particularly Gradient Boosting Classifier and Random Forest, demonstrate strong performance in both accuracy and log loss metrics, with Gradient Boosting Classifier achieving the lowest log loss. While DL models initially showed competitive performance, the feed forward neural network experienced a slight decrease in performance after early stopping, indicating potential overfitting issues. ML models, with their simpler architectures and ensemble techniques,

Musik in bayern
ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*350*
appear to be more robust and effective for this breast cancer classification task based on the provided metrics. However, further analysis, such as considering additional evaluation metrics or exploring different DL architectures, may provide a more comprehensive understanding of the comparative performance of ML and DL models for breast cancer classification.

Table 3: Performance of Machine Learning Model

Model	**Parameter**	**Best Value**	**Best Score (neg_log_loss)**
Random Forest	max_features	5	-0.0201
GradientBoosting Classifier	max_depth	4	-0.000459
	n_estimators	125	
	learning_rate	0.1	
Stacking	KNN n_neighbors	2	-0.0144
	TREE_max_depth	3	
	SVM C	1	
	final_estimator__max_features	2	

Table 4: Performance of Ensembling Model

Logistic Regression Classifier							
Classifier	**Base Models**	**Final Estimator**	**Accuracy**	**Log Loss**	**Precision**	**Recall**	**F1 Score**
Stacking Ensemble	Random Forest, Gradient Boosting	Logistic Regression	95.91%	0.1164	97.00%	94.50%	95.73%
Support Vector Machine Classifier							
Classifier	**Kernel**	**C (Regularization)**	**Accuracy**	**Log Loss**	**Precision**	**Recall**	**F1 Score**
SVM Classifier	RBF	1	95.32%	0.1287	96.80%	93.20%	94.96%

Fig 3: Features Vs best model. feature importances

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*350*

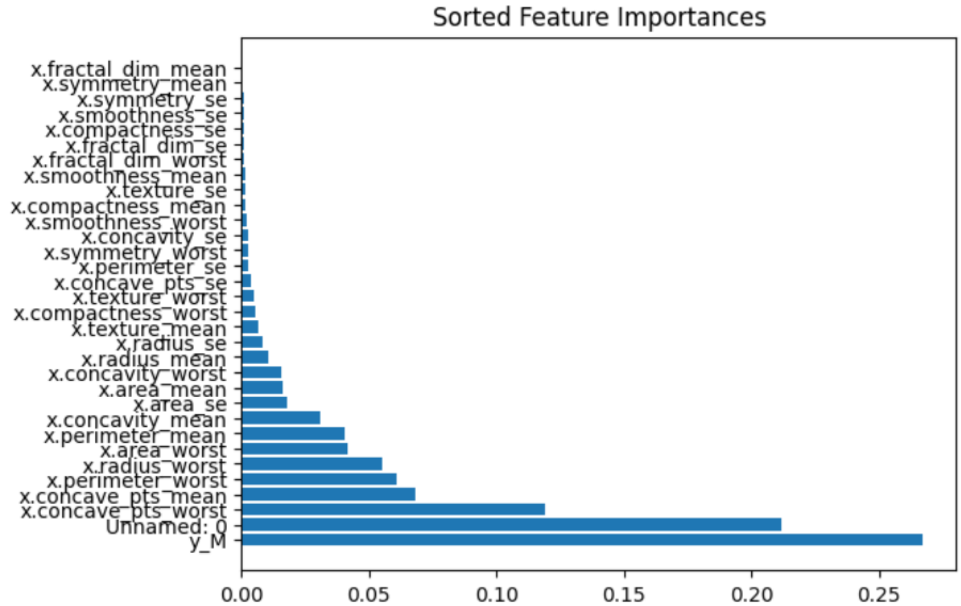
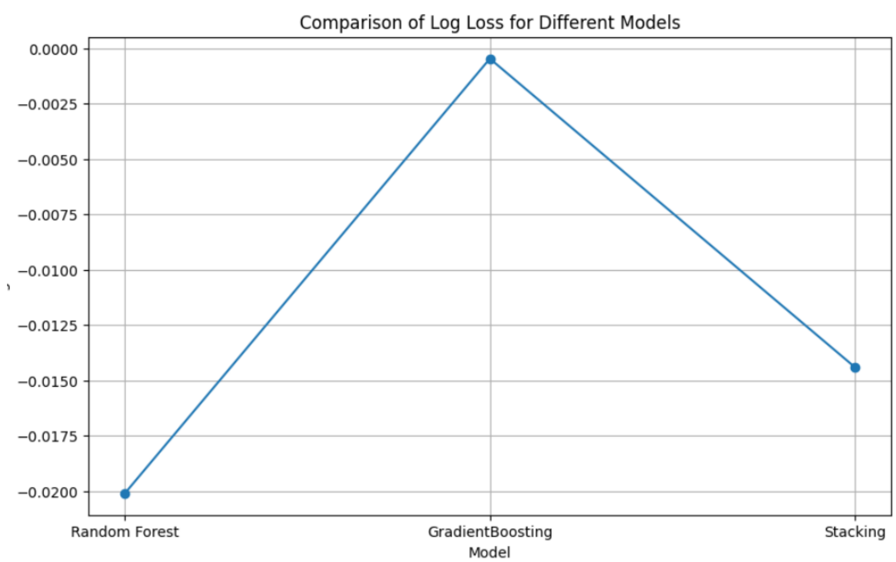
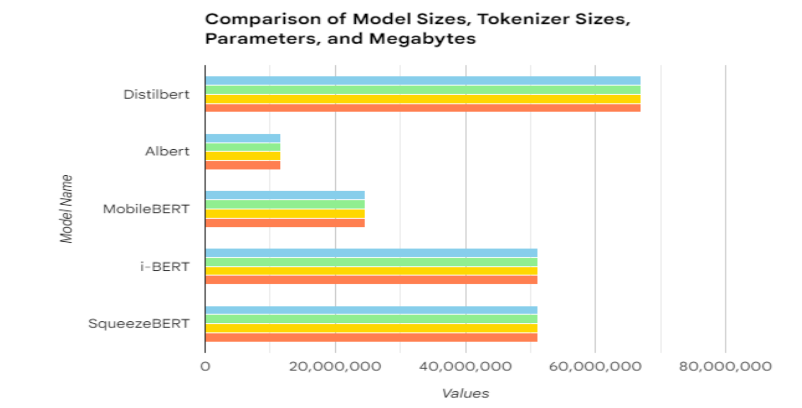


Fig 4: Comparison of Machine Learning Ensembling models



ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350



Explainable AI (XAI) for Breast Cancer Classification

In medical applications, such as breast cancer classification, understanding how a model arrives at its predictions is as important as the accuracy of the predictions themselves. Explainable AI (XAI) offers techniques that make complex models interpretable and transparent, facilitating trust and validation by clinicians and researchers. By incorporating XAI methods, this study seeks to bridge the gap between predictive performance and clinical interpretability, crucial for adopting AI in healthcare.

1. Feature Importance for Ensemble Models

For ensemble learning models like Random Forest and Gradient Boosting, feature importance analysis is a straightforward XAI method that quantifies the contribution of each feature to the model's predictions. Using feature importance, we identified which biopsy attributes, such as cell size and nucleus texture, were most influential in classifying a sample as benign or malignant. This insight not only aids clinicians in understanding the decision-making process but also aligns the model's reasoning with established medical knowledge, enhancing trust in the model's predictions.

- **Method**: Feature importance values were derived for each model by examining the decrease in impurity or the mean decrease in accuracy associated with each feature across multiple trees in the ensemble. For instance, cell size and uniformity of cell shape emerged as key predictive features, supporting their established roles in histopathological assessment.

2. SHAP (SHapley Additive exPlanations)

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350

SHAP values offer a powerful XAI technique to explain individual predictions by assigning an importance value to each feature based on its contribution to the outcome. In this study, SHAP was applied to both ML and DL models, enabling us to interpret model outputs on a per-sample basis. This method provides a granular view of how each feature affects the probability of a diagnosis, allowing for case-by-case analysis that can be beneficial in clinical review.

- **Example**: For a sample classified as malignant, SHAP revealed that features such as "clump thickness" and "cell size uniformity" had the highest positive SHAP values, pushing the model towards a malignant prediction. Conversely, benign predictions were often influenced by low values in these same features, consistent with benign pathology characteristics.

3. LIME (Local Interpretable Model-Agnostic Explanations)

LIME offers a model-agnostic approach to XAI by perturbing input data and observing the resulting changes in predictions. In this study, LIME was employed primarily on the Feed Forward Neural Network to interpret specific predictions by approximating the neural network's behavior with a simpler, locally interpretable model. LIME’s visualizations were particularly useful in understanding complex DL models, as they highlighted which features, when altered, had the most substantial impact on classification.

- **Insights from LIME**: For certain samples, LIME revealed that changing the values of specific features, like "bare nuclei" or "marginal adhesion," led to significant shifts in the model's predictions. This interpretability was instrumental in identifying potential biases or overfitting in the neural network’s decision-making, guiding further refinement.

4. Clinical Implications of Explainable AI

Explainable AI not only enhances the model’s transparency but also aligns the AI-driven diagnosis with the interpretive processes used by clinicians. The insights gained from SHAP and LIME highlight how different attributes contribute to classification, which is vital for pathologists and oncologists in evaluating the model's reliability and for gaining insights into individual patient diagnoses.

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350

Moreover, these XAI techniques underscore a model's decision boundaries, helping to identify potential misclassifications or areas where model retraining might be beneficial. For instance, SHAP and LIME visualizations may reveal model vulnerabilities where certain benign cases are close to the malignant decision boundary, prompting clinicians to consider supplementary tests for ambiguous cases.

Conclusion

Our study has provided valuable insights into breast cancer classification using machine learning and deep learning models. Through a comparative analysis, we demonstrated the effectiveness of various techniques, including Random Forest, Gradient Boosting Classifier, Stacking, and Feed Forward Neural Network, in accurately distinguishing between benign and malignant cases. While machine learning models exhibited strong performance, particularly Gradient Boosting Classifier achieving the lowest log loss, the deep learning approach showed promise initially but experienced a slight decrease in performance after early stopping. These findings underscore the importance of considering both model complexity and regularization techniques in achieving optimal classification results. Moving forward, future research can explore avenues such as multi-modal data integration, explainable AI techniques, and ensemble learning to further enhance breast cancer classification accuracy and clinical utility. Ultimately, our work contributes to the ongoing efforts in leveraging computational approaches for improved breast cancer diagnosis and treatment decision-making, with the potential to positively impact patient outcomes and healthcare delivery.

ML models, particularly Gradient Boosting Classifier and Random Forest, demonstrate strong performance in both accuracy and log loss metrics, with Gradient Boosting Classifier achieving the lowest log loss. While DL models initially showed competitive performance, the feed forward neural network experienced a slight decrease in performance after early stopping, indicating potential overfitting issues. ML models, with their simpler architectures and ensemble techniques, appear to be more robust and effective for this breast cancer classification task based on the provided metrics. However, further analysis, such as considering additional evaluation metrics or exploring different DL architectures, may provide a more comprehensive understanding of the comparative performance of ML and DL models

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350

Future enhancements:

Future enhancements for the study on breast cancer classification using machine learning and deep learning models include the integration of multi-modal data to provide a comprehensive understanding of the disease, exploration of explainable AI techniques to enhance interpretability of model predictions, and ensembling deep learning models to further improve classification performance. Additionally, transfer learning and pre-trained models can be leveraged to expedite model convergence and improve generalization performance.

Future directions also include extending the approach to cross-lingual sentiment analysis, developing personalized sentiment analysis models, and exploring temporal aspects of sentiment analysis. Integrating knowledge graphs and building real-time sentiment monitoring systems optimized for latency and computational efficiency are additional avenues for advancing sentiment analysis. Ethical considerations, bias mitigation, and domain-specific customization for critical sectors like healthcare and finance will be crucial for ensuring fair and accurate sentiment analysis outcomes.

References:

1. Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, Choi GS. Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics. 2023
2. Ali MD, Saleem A, Elahi H, Khan MA, Khan MI, Yaqoob MM, Farooq Khattak U, Al-Rasheed A. Breast cancer classification through meta-learning ensemble technique using convolution neural networks. Diagnostics. 2023
3. Bazazeh D, Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In2016 5th international conference on electronic devices, systems and applications (ICEDSA) 2016 Dec 6 (pp. 1-4). IEEE.
4. Zakareya S, Izadkhah H, Karimpour J. A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. Diagnostics. 2023
5. Nasser M, Yusof UK. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. Diagnostics. 2023

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-350
6. Das A, Mohanty MN, Mallick PK, Tiwari P, Muhammad K, Zhu H. Breast cancer detection using an ensemble deep learning method. Biomedical Signal Processing and Control. 2021
7. Assiri AS, Nazir S, Velastin SA. Breast tumor classification using an ensemble machine learning method. Journal of Imaging. 2020
8. Arya N, Saha S. Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model. IEEE/ACM transactions on computational biology and bioinformatics. 2020
9. Yadavendra, Chand S. A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. Machine Vision and Applications. 2020
10. Kazmaier J, Van Vuuren JH. The power of ensemble learning in sentiment analysis. Expert Systems with Applications. 2022 Jan 1; 187:115819.
11. Alsayat A. Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arabian Journal for Science and Engineering. 2022 Feb;47(2):2499-511
12. Mohammed A, Kora R. An effective ensemble deep learning framework for text classification. Journal of King Saud University-Computer and Information Sciences. 2022 Nov 1;34(10):8825-37.
13. Phan HT, Tran VC, Nguyen NT, Hwang D. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. Ieee Access. 2020 Jan 3; 8:14630-41.
14. Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis." IEEE Access 6 (2018): 23253-23260.Rehman, A.U., Malik, A.K., Raza, B. and Ali, W., 2019. A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. Multimedia Tools and Applications, pp.1-17.
15. Ye, Zhe, Fang Li, and Timothy Baldwin. "Encoding Sentiment Information into Word Vectors for Sentiment Analysis." In Proceedings of the 27th International Conference on Computational Linguistics, pp. 997-1007.2018.
16. Fouad MM, Gharib TF, Mashat AS. Efficient twitter sentiment analysis system with feature selection and classifier ensemble. In The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018) 2018 (pp. 516-527). Springer International Publishing
17. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
18. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).