TRANSFORMING TABULAR DATA FOR MULTI-MODALITY: ENHANCING BREAST CANCER METASTASIS PREDICTION THROUGH DATA CONVERSION

Faseela Abdullakutty [1], Younes Akbari [1], Somaya Al-Maadeed [1], Ahmed Bouridane [2], Rifat Hamoudi [2]

[1]Qatar University, Qatar,
[2]University of Sharjah, United Arab Emirates

ABSTRACT

Breast cancer metastasis prediction plays a key role in clinical decision-making and secondary analysis. Traditionally, metastasis classification models have been developed using structured tabular clinical data, but these approaches may result in data loss and lack of contextual information. A multi-modal approach is presented in this article for predicting breast cancer metastasis by converting structured clinical data into unstructured text, which provides more contextual information, and then converting that text into histopathology images. For text classification, features were extracted and fine-tuned. Using Logistic Regression and XGBoost classifiers, these extracted features exhibited enhanced performance. The accuracy of the metastasis detection was further enhanced by fine-tuning. A pre-trained diffusion model was used to generate histopathology images based on the same clinical data to address the multimodality gap. The classification of the features extracted from these images, using pre-trained vision models like VGG-16 and ViT, provided similar results to traditional tabular predictions. A multi-modal early fusion approach was then created by combining vision-derived features with text-based features from the BERT. Using unstructured text and histopathology images can effectively address multi-modal data limitations, providing a promising alternative for future research and providing a context-rich approach to breast cancer metastasis prediction.

Index Terms— Breast cancer metastasis, data transformation, multi-modality

1. INTRODUCTION

Breast cancer remains one of the most prevalent cancers across the globe, with early detection playing a crucial role in improving survival rates [1]. To achieve a successful cancer treatment, it is important to know that the spread of cancer from the primary site to other parts of the body is known as metastasis. To give patients a good prognosis and guide treatment decisions, breast cancer metastases must be diagnosed accurately to determine their prognosis [2].

It has become increasingly common for machine learning (ML) techniques to be used in medical diagnosis as they have the capability of analyzing large datasets and identifying complex patterns. A traditional approach to the detection of breast cancer metastases has relied on the use of structured clinical data, which typically consists of numerical values, categorical information, and coded entries. There is, however, a potential problem with this structured approach due to its rigid format. Because of this, it could miss crucial context and impact the accuracy of cancer diagnosis [3]. The use of transformer-based models, like BERT (Bidirectional Encoder Representations from Transformers) [4], offers a more flexible approach to data analysis, especially when dealing with unstructured clinical text, which has the advantage of providing better results. BERT-based models are highly efficient at extracting context-rich information from text data, thus allowing more nuanced analysis and improved accuracy when analyzing text data. It has been demonstrated that unstructured text from medical reports, clinical notes, and other sources can be analyzed using transformer models to enhance cancer detection, including the detection of metastatic breast cancer [5].

By converting structured clinical data into unstructured text data and subsequent histopathology images, the present study investigated an innovative method for detecting breast cancer metastasis. To predict metastasis, the unstructured clinical text was analyzed and classified using transformer-based techniques. Through the use of pre-trained diffusion models [6], this study illustrates a process for transforming tabular clinical data into text that can then be used to generate corresponding histopathology images. In this multi-modal data transformation approach, we strive to utilize the efficiency of structured data processing while incorporating the contextual depth of unstructured text and the visual interpretation provided by histopathology images. Figure. 1 shows the metastasis prediction using different data modalities and multi-modal early fusion using generated images and text data.

It is possible to create a more comprehensive view of medical cases by converting tabular clinical data into text, thereby reducing the risk of data loss and allowing for the addition of detailed narratives [3]. Through the use of BERT-based text classification models, we demonstrate improved performance in the detection of breast cancer metastases. Addition-

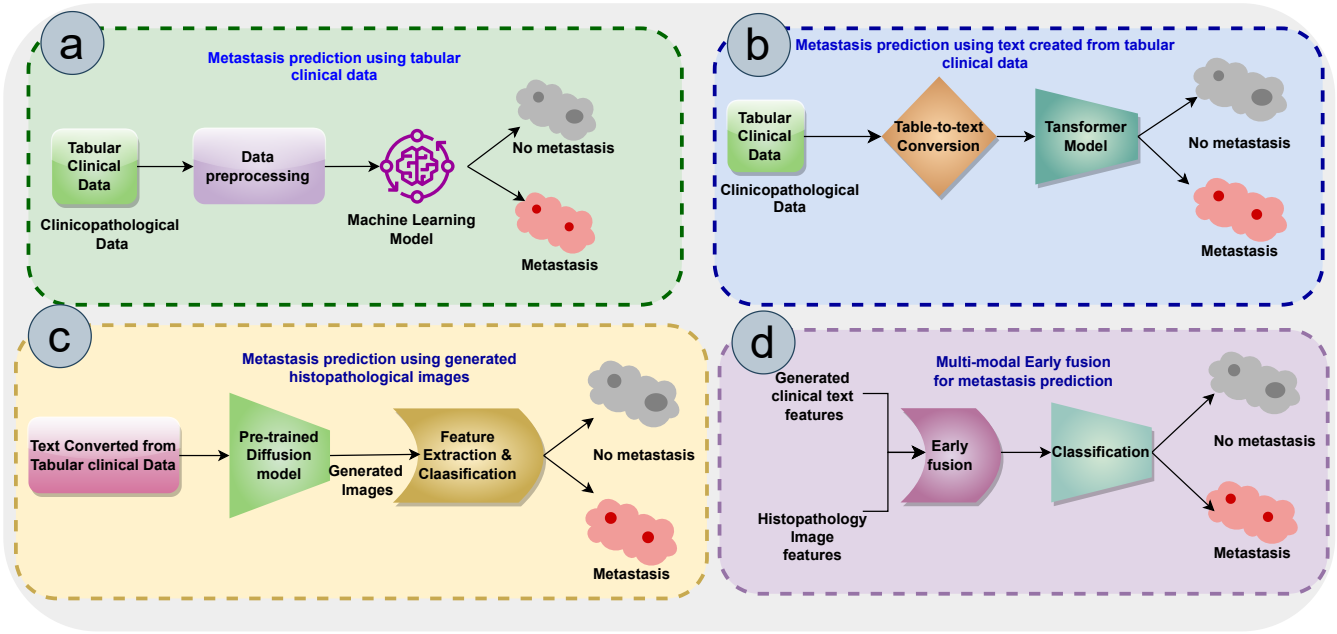

Fig. 1: Predicting Breast Cancer Metastasis using different data modalities. a) Tabular data, b) converted text data, c) generated histopathology images, and d) multi-modal early fusion using image and text features.

ally, we generate histopathology images from these textual representations, which offers a new perspective on the data for enhancing cancer detection accuracy. The use of diffusion models to create images from text-based data provides a unique avenue for incorporating multi-modal analysis into the diagnostic process.

This combination of structured data, unstructured text, and generated images creates a more robust framework for breast cancer metastasis detection, offering greater flexibility and adaptability in clinical scenarios. VGG-16 [7] and ViT [8] models were used to extract vision features from generated images. These features were then combined with text features extracted using the BERT model to form early fusion in a multi-modal context. The experiments used a metastasis dataset [9], which has 20 features. The dataset has information on patient profile, treatment history, tumour information and pathology information. The key contributions of this research are as follows:

- Multi-Modal Data Transformation for Breast Cancer Metastasis Prediction.
- Comparative Analysis of Prediction Models Across Different Modalities: A comprehensive comparison of different methodologies for predicting breast cancer metastasis using structured data, unstructured text, generated images and multi-modal data.
- Evaluation of Multi-Modal Early Fusion with Pre-Trained Models: An assessment of the effectiveness of multi-modal early fusion techniques using features extracted from pre-trained transformer and diffusion models.

2. RELATED WORKS

Predicting breast cancer metastasis is essential for improving prognosis, treatment planning, early intervention, and resource allocation. Metastasis, the spread of cancer from its original site to other parts of the body, significantly impacts cancer outcomes. Accurate predictions enable clinicians to tailor treatment plans, potentially reducing the spread and improving patient outcomes. Machine learning (ML) is crucial in this process, as it analyzes large datasets to uncover complex patterns and relationships that are not immediately apparent to human experts. ML facilitates risk stratification, helping identify high-risk patients, supports personalized medicine by predicting treatment responses, and automates tasks like analyzing medical images or genetic data, leading to faster decision-making and reduced workloads for medical professionals. Continuous learning in ML models ensures they remain accurate over time. This study focuses on early-stage breast cancer treatment, aiming to reduce unnecessary surgeries, and explores methods involving clinical, textual, and image data.

The authors of [10] proposed a method that combined tumour detection, segmentation, and metastatic prediction to improve clinical outcomes. Mask R-CNN was used to identify and segment tumours in ultrasound images, extract-

ing peritumoral tissue for cancer spread assessment. A deep learning model was then used to predict axillary lymph node (ALN) metastasis, a common site for breast cancer spread. The CNN-based Computer-Aided Prognosis (CAP) system enhances ALN metastasis prediction accuracy, providing more reliable results for early-stage breast cancer. This approach could significantly impact clinical decision-making and could lead to fewer unnecessary surgeries and postoperative complications. The authors of [11] developed a Natural Language Processing algorithm to extract key features about breast cancer from medical records. It also used machine learning algorithms to predict patient recurrence of breast cancer. The TF-IDF method was used to identify the importance of each term in reports. A structured medical library was created using the NLP algorithm. Machine learning algorithms like J48, NaiveBayes, and SVM were used to predict factors influencing breast cancer recurrence.

Using F-18 fluorodeoxyglucose positron emission tomography/computed tomography (PET/CT), the authors of [12] presented a method to predict axillary lymph-node (ALN) metastasis in invasive ductal breast cancer (IDC). The Volume of Interests (VOIs) was drawn from the PET scan using a 3D slicer. The Pyradiomics package extracted 792 radiomic features per patient. A radiomics prediction model was developed for 75 patients and validated in 25 patients using the XGBoost algorithm. To evaluate the effectiveness of the model in predicting ALN metastasis in patients with IDC, the sensitivity, specificity, and accuracy of the model were evaluated. This alleviated data limitation issues. The paper [13] presented four machine learning models developed for male breast cancer patients using data from the SEER database and a hospital between 2010 and 2020. The models included extreme gradient boosting, K Nearest Neighbour, decision tree, and support vector machine. A 7:3 training and testing ratio was used, and unbalanced data was addressed using SMOTE resampling. The hyperparameters of the models were optimized using a grid search method. The SHapley Additive exPlanations framework was used to interpret the optimal model, providing global and local interpretability plots.

The model [14] aimed to quickly and accurately find segment lesion regions in high-resolution breast cancer pathology sections. It consisted of three main modules: Image Preprocessing, ROI Recognition, and Lesion Region Segmentation. The model was evaluated on a breast cancer dataset from two medical institutions, achieving an average recognition precision of 0.936 for the region of interest and an F1-score of 0.787, outperforming similar works in terms of speed and precision.

Natural Language Processing (NLP) is a crucial tool in the medical field, utilizing Machine Learning and Deep Learning techniques to extract valuable insights from Electronic Health Records (EHR) [15]. It is used in various healthcare applications, such as cardiovascular risk factor detection, heart condition identification, oral disease diagnosis, and cancer tumour detection from radiology images. NLP techniques, such as entity recognition, are essential for processing medical text data and supporting decision-making processes in healthcare. Despite challenges posed by unique clinical language and idioms, NLP techniques are instrumental in advancing healthcare information research and improving the efficiency of managing clinical notes in the healthcare industry. CancerBERT [16] models are a significant advancement in extracting breast cancer phenotypes from clinical texts. They were built on transformer-based models like BERT and applied specialized domain knowledge to improve accuracy. The vocabulary of the models was customised based on relevant terms within the breast cancer domain, allowing it to better understand and interpret clinical language. The research focused on eight key breast cancer phenotypes, such as hormone receptor type, tumour size, and cancer grade, to support clinical decision-making and potentially impact patient treatment plans. CancerBERT models outperformed other baseline models, demonstrating their effectiveness in extracting breast cancer phenotypes from complex clinical texts.

GANs are increasingly used in medical imaging, particularly in cross-modality image synthesis and MRI imaging. However, challenges include using generated images for clinical diagnosis and decision-making, requiring valid metrics, and addressing data fidelity loss during image translation. Losses in medical image generation using GANs include shape, perceptual, structural, style-content, self-regulation, data fidelity, and regulation [17]. Diffusion models in medical imaging have gained interest due to their potential in tasks, like anomaly detection, segmentation, denoising, classification, and reconstruction. Existing surveys have mainly focused on deep generative models, leaving a research gap. Diffusion models offer an alternative to traditional generative models like VAEs and GANs, showing promise in image generation, segmentation, and anomaly detection [18]. The study [19] presents a novel approach using diffusion probabilistic models to generate high-quality histopathology images of brain cancer. The method aims to produce a diverse array of complex morphologic features in histopathology images, which can be more complex than typical image sets. The end-to-end process employs diffusion probabilistic models specifically designed for Hematoxylin and Eosin (HE) stained histopathology images, demonstrating superior performance compared to conventional models like proGAN. The core of the method involves training the model with score matching losses to achieve the desired image synthesis. This approach aids in better diagnosis and understanding of brain cancer. However, the current method faces challenges, particularly in sampling time. Future work could focus on optimizing the model to reduce sampling time, improve efficiency and accessibility. This study contributes to medical imaging by offering a new avenue for generating detailed and diverse histopathology images.

3. METHOD

In this study, we investigated three approaches for breast cancer metastasis detection: structured data classification, BERT-based text data classification, and image generation and classification. The following steps outline the methodology for this experiment:

3.1. Dataset

The clinical dataset [9] for detecting breast cancer metastasis has been used for the experiments described in this article. Among the features included in this tabular clinical pathology data are age, nulliparity, oral contraception, menopause, breast cancer history in the family, number of full-term pregnancies, obesity, estrogen receptor status, progesterone receptor status, tumour size, histology, vascular invasion, tumour grade, surgery, adjuvant chemotherapy, radiotherapy, trastuzumab, and hormone therapy. To make the structured data suitable for machine learning classification algorithms, it was processed to convert it into numerical values. In addition, a text-based version of this dataset was compiled. Every record was transformed into a structured text format with key information encoded as sentences or paragraphs.

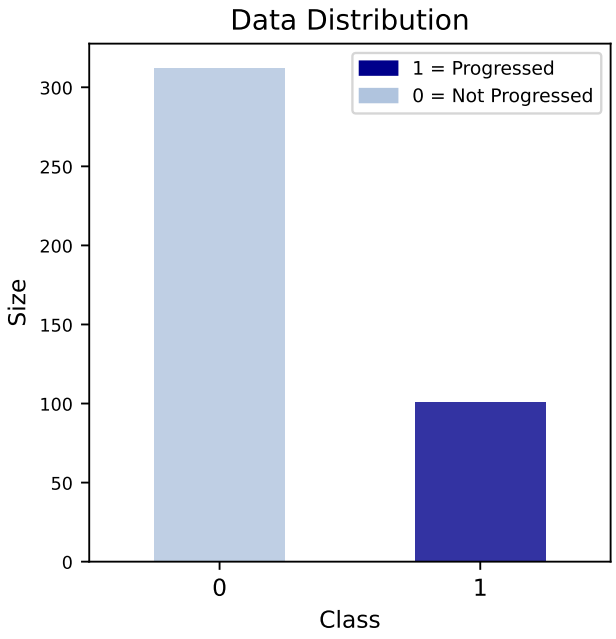

Fig. 2: The data distribution of Metastasis dataset [9]

Figure. 2 illustrates the data distribution of the metastasis dataset, which has two classes. Label 1 indicates a progressed state, whereas label 0 represents no progression scenario. This dataset after cleaning and removing unknown labels and values, has a train set of size 330 and a test set of size 83 with 20 features.

3.2. Metastasis prediction

This article discusses data transformation to support metastasis prediction, specifically through the use of tabular clinical pathology data. The tabular data was employed to predict metastasis, consistent with the approach described in [9], using machine learning classifiers. These predictions served as a baseline. The structured tabular data was then transformed into unstructured text, which was subsequently used as part of the prompt in pre-trained diffusion models to generate histopathology images. A detailed explanation of the various steps in the methodology outlined in this article is provided in the following sections.

3.2.1. Structured Data classification

The process of structured data classification entailed the application of various machine learning algorithms, including Logistic Regression(LR), Decision Trees (DT), K-Nearest Neighbors (KNN), and XGBoost, to numerical datasets. The dataset was split into two subsets: training and testing, typically with an 80/20 ratio. The machine learning models were trained on the training subset and subsequently evaluated on the testing subset. The accuracy metrics were used to assess and compare the performance of the different classification algorithms. Figure 1 (a) illustrates the prediction task derived from structured tabular data.

3.2.2. Unstructured clinical text classification

The tabular data was converted to unstructured text data. This text data was then used to predict metastasis progress. In text classification, both feature extraction BERT model and fine-tuning of the BERT model were carried out. The extracted features were classified using LR and XGBoost to compare the performance with the baseline. BERT and Bio ClinicalBERT were also fine-tuned with the unstructured text data for prediction. This process is shown in the Figure. 1 (b).

The structured text data was tokenized and formatted to meet BERT's input requirements, and the same 80/20 train-test split ratio was used. Fine-tuning involved adjusting the weights of a pre-trained BERT model to adapt it for the task of breast cancer metastasis detection. Similar to baseline Accuracy was reported for the comparison of performance.

3.2.3. Image Generation and Classification

To utilise the visual cues as another modality for precision, histopathology images were generated using a pre-trained diffusion model, Prompt2MedImage [6]. This text-to-image diffusion model can generate high-quality medical images from textual prompts. It relies on a fixed, pre-trained text encoder (specifically, the CLIP ViT-L/14 model), with standard settings. The converted text which has a tumour description and histology image type is given as a part of the prompt as shown

in the Figure. 3. The prompt input was " Show histopathology images stained with Hematoxylin and Eosin, captured with 40X magnification, aligned with the metastasis description: ***Text*** ". Here in the prompt, ***Text*** indicates the text generated using tabular data. Based on this histopathology image was created. Such generated images are shown below in Figure. 4.

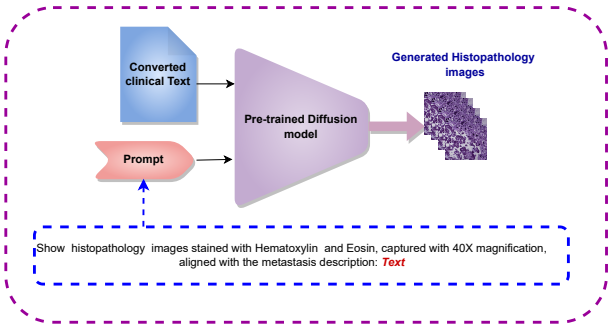

Fig. 3: Generating images from converted clinical text

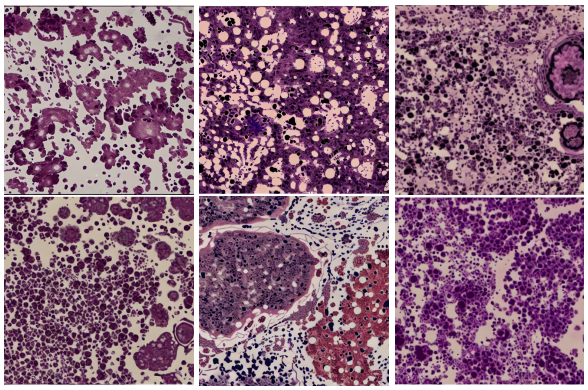

Fig. 4: Generated histopathology image sample

For a comparative analysis with baseline and text classification results, pre-trained vision models, including VGG-16 and ViT we reused to extract features from these images. Extracted features were classified using ML classifiers including LR and XGBoost. Figure. 1 (c) shows the entire process of converting the table to the image to predict metastasis.

3.2.4. Metastasis prediction using multi-modal data

To predict metastasis and assess the impact of multi-modal data, early fusion based on concatenation was explored using text and image data derived from clinical tabular information. The text and image features were extracted from the data using pre-trained models, including BERT, VGG-16, and Vision Transformer (ViT). These extracted features were concatenated and classified using various machine-learning classifiers. An illustration of this method is shown in Figure. 1 (d).

4. RESULTS

The tabular data, converted text and generated histopathology images were used with different classifiers and pre-trained models to predict metastasis. The results are shown in Table. 1. The BERT model when fine-tuned with the converted texts performed the best compared to other scenarios. The BERT performance is 86%, which is 10% higher than the baseline model [9].

Table 1: Performance comparison on metastasis prediction using tabular data for the generated text and generated image data.

Model	Data Type	Accuracy
Logistic Regression	Tabular Data	76%
Decision Tree	Tabular Data	71%
KNN	Tabular Data	75%
BERT	Converted Clinical Text	**86%**
Bio-ClinicalBERT	Converted Clinical Text	80%
ViT+ Logistic Regression	Images generated from converted clinical text	73.49%
BERT+XGBoost	Converted Clinical Text	77.10%
BERT+Logistic Regression	Converted Clinical Text	78.31%
BERT+ViT+XGBoost	Text, Image	77.11%

Features extracted from the VGG-16 and Vision Transformer (ViT) architectures were subjected to classification using Logistic Regression (LR) and XGBoost algorithms. The outcomes demonstrated that ViT-derived features yielded comparable classification results with both the LR and XGBoost classifiers. However, a distinct pattern emerged when using features extracted from VGG-16: Logistic Regression achieved a prediction accuracy of 55.42%, while XGBoost significantly outperformed it with a classification accuracy of 71.08%. This noticeable increase in accuracy when using the XGBoost classifier suggests a potential synergy between its tree-based approach and the structure of VGG-16-derived features, highlighting the importance of classifier choice in achieving optimal performance.

In the early fusion approach, combining features from BERT and Vision Transformer (ViT) and then using XGBoost for classification yielded an accuracy of 77.11%. When these same features were classified with Logistic Regression (LR), the accuracy slightly improved to 78.31%. However, when the features from VGG-16 were combined with those from BERT and classified with the same methods, the accuracy significantly decreased to 54.22%.

4.1. Discussion

The experiments described in this article explored the application of models like BERT for text classification in digital pathology and investigated image generation using diffusion models. Text-based data derived from structured clinical tables can contain additional context and details that may

be crucial for diagnosis. By converting tables to text, digital pathology systems can incorporate this context into their analyses, leading to more accurate and nuanced diagnostic insights.

It is evident from Table. 1 that training the BERT model on converted text data improved overall performance. Despite Bio-ClinicalBERT's lower performance compared to BERT, even after training with the dataset used for evaluation, it still performed better than tabular data-based and pre-trained text and image feature-based prediction using ML classifiers. This suggests that to achieve optimal performance, models need fine-tuning with task-specific data.

Regarding image generation, experiments in this article used a pre-trained model trained on a variety of radiology images rather than histopathology images. Although near-realistic histopathology images were generated with this method, fine-tuning the pre-trained model with histopathology images from public multi-modal datasets with histopathology images and texts can lead to improved image generation results.

The multi-modal early fusion approach achieved an accuracy of 77.11% when combining features from BERT and Vision Transformer, slightly improved by Logistic Regression to 78.31%, but decreased sharply to 54.22% when combining VGG-16 and BERT features. This difference in results could be due to several factors. Pre-trained feature extraction models like BERT and ViT are designed for specific types of data; BERT is optimized for text, while ViT and VGG-16 are tailored to process image data. The observed accuracy drop when using VGG-16 with BERT may suggest that the combination of these specific features lacks the synergy observed with other pairings. VGG-16, being an older architecture focused on image data, might not extract as nuanced or complementary features as ViT when paired with BERT's text-based features.

Additionally, the data's modality and inherent properties might play a role. VGG-16, designed for object recognition tasks, may not capture the same depth of information in complex medical images as ViT. This, combined with BERT's focus on linguistic structures, could lead to a mismatch in feature extraction and ultimately lower accuracy. Overall, these results underscore the importance of considering the compatibility of different feature extraction models when performing multi-modal analysis. It also highlights the need for careful experimentation with different classifier combinations to determine the most effective approach for a given dataset and task.

As can be seen from Figure. 2, there is an imbalance between the classes in the dataset. To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning can be utilized to balance the dataset and enhance the focus on the minority class. The accuracy of the model can be further enhanced by adjusting the weights of the minority classes in the loss function. In addition to the standard accuracy metric, sensitivity and specificity will also be used to evaluate model performance. A high-sensitivity approach is used in this approach to identify metastatic cases with high specificity to identify non-metastatic cases with high accuracy. A robust and clinically relevant model for predicting breast cancer metastasis will be developed as a result of these measures.

This approach reduced data loss and enhanced information retention. Structured data can sometimes lead to the loss of important information due to rigid data structures. This experiment also opened new avenues for advanced research and development, fostering innovation in diagnostic tools and smarter pathology systems. By combining advanced models with creative data transformation techniques, this study contributes to the evolution of digital pathology.

5. CONCLUSION

This study demonstrated the effectiveness of a multi-modal data transformation approach in breast cancer metastasis prediction. By transforming structured clinical data into unstructured text and histopathology images, the accuracy of breast cancer metastasis prediction can be significantly improved. This approach leverages the contextual richness of unstructured text and the visual insights provided by generated histopathology images. The use of BERT-based transformer models for text classification resulted in an improvement in metastasis detection performance, highlighting the potential of these advanced models for analyzing clinical texts. Diffusion models for image generation added a new dimension to analysis by generating histopathology images from text-based data. This technique bridges the multi-modality gap and provides additional insights for accurate breast cancer metastasis prediction. This study paves the way for new research opportunities in breast cancer metastasis prediction by demonstrating how multi-modal data can be generated from existing sources, potentially addressing the shortage of multi-modal datasets. Through this approach, it was intended to enhance the detection performance while alleviating data limitation issues.

Acknowledgment

The research reported in this publication was supported by the Qatar Research Development and Innovation Council [ARG01-0513-230141].

6. REFERENCES

[1] Ziquan Zhu, Shui-Hua Wang, and Yu-Dong Zhang, "A survey of convolutional neural network in breast cancer," *Computer modeling in engineering & sciences: CMES*, vol. 136, no. 3, pp. 2127, 2023.

[2] Zexian Zeng, Liang Yao, Ankita Roy, Xiaoyu Li, Sasa Espino, Susan E Clare, Seema A Khan, and Yuan Luo, "Identifying breast cancer distant recurrences from electronic health records using machine learning," *Journal of healthcare informatics research*, vol. 3, pp. 283–299, 2019.

[3] Chérubin Mugisha and Incheon Paik, "Bridging the gap between medical tabular data and nlp predictive models: A fuzzy-logic-based textualization approach," *Electronics*, vol. 12, no. 8, pp. 1848, 2023.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] Santiago González-Carvajal and Eduardo C Garrido-Merchán, "Comparing bert against traditional machine learning text classification," *arXiv preprint arXiv:2005.13012*, 2020.

[6] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich, "Radiology objects in context (roco): a multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer, 2018, pp. 180–189.

[7] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] Ikram Maouche, Labib Sadek Terrissa, Karima Benmohammed, and Noureddine Zerhouni, "An explainable ai approach for breast cancer metastasis prediction based on clinicopathological data," *IEEE Transactions on Biomedical Engineering*, 2023.

[10] Yan-Wei Lee, Chiun-Sheng Huang, Chung-Chih Shih, and Ruey-Feng Chang, "Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks," *Computers in Biology and Medicine*, vol. 130, pp. 104206, 2021.

[11] Amal Alzu'bi, Hassan Najadat, Wesam Doulat, Osama Al-Shari, and Leming Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13787–13800, 2021.

[12] Bong-Il Song, "A machine learning-based radiomics model for the prediction of axillary lymph-node metastasis in breast cancer," *Breast Cancer*, vol. 28, pp. 664–671, 2021.

[13] Xuhai Zhao and Cong Jiang, "The prediction of distant metastasis risk for male breast cancer patients based on an interpretable machine learning model," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, pp. 74, 2023.

[14] Yuanyue Lu, Jun Zhang, Xueyu Liu, Zhihong Zhang, Wangxing Li, Xiaoshuang Zhou, and Rongshan Li, "Prediction of breast cancer metastasis by deep learning pathology," *IET Image Processing*, vol. 17, no. 2, pp. 533–543, 2023.

[15] Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner, "Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review," *Computers in Biology and Medicine*, vol. 155, pp. 106649, 2023.

[16] Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang, "Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records," *Journal of the American Medical Informatics Association*, vol. 29, no. 7, pp. 1208–1216, 2022.

[17] Nripendra Kumar Singh and Khalid Raza, "Medical image generation using generative adversarial networks: A review," *Health informatics: A computational perspective in healthcare*, pp. 77–96, 2021.

[18] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, p. 102846, 2023.

[19] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati, "A morphology focused diffusion probabilistic model for synthesis of histopathology images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2000–2009.