

Classifying gastric cancer carcinoma stages with deep semantic features and GLCM texture features

Sikandar Ali *, Samman Fatima *, Ali Hussain *, Maisam Ali*, Muhammad Yaseen*, Tagne Poupi Theodore Armand*, Hee-Cheol Kim**

* Dept. of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

** College of AI Convergence, Institute of Digital Anti-Aging Healthcare, u-AHRC, Inje University, Gimhae 50834, Korea
sikandarshigri77@gmail.com, samman.1511@gmail.com, alihussainnrana@gmail.com, maisamali053@gmail.com, shigriyaseen@gmail.com, poupiarmand2@gmail.com, heeki@inje.ac.kr

Abstract— Gastric cancer is one of the leading health issues that contributes to cancer related deaths. The tricky thing about cancer is that it often goes undetected until at higher stages, which makes treatment less effective. The significant death rate from gastric cancer highlights the importance of a precise and prompt diagnosis. This paper aims to tackle this problem by proposing an approach to classify the early and advanced stages of gastric cancer. This importance of this study stems from its two-pronged strategy, which provides a deeper understanding of stomach cancer stages using texture analysis and deep learning. We take advantage of the strengths of deep learning features, Gray Level Co-occurrence Matrix (GLCM) features, and machine learning algorithm to create a diagnostic tool that is more precise and accurate. Medical images from gastric cancer dataset showing early and advanced stages of gastric cancers carcinoma are included to develop this model. Our method combines the effectiveness of texture features extracted from GLCM combined with deep semantic features and classify the stages with machine learning model. We carefully evaluated Machine learning classifiers namely Support Vector Machine (SVM), Decision Tree (DT), and K-nearest neighbour (KNN) to classify the early and advanced stages. Each classifier was evaluated with different performance measures. The Support Vector Machine (SVM) classifier demonstrated the best performance with an accuracy of 96.93%. This highlights the potential of SVM for diagnosing different cancer stages, which could have positive implications, for clinical practice.

Keywords— Gastric Cancer, Machine Learning, Support Vector Machine (SVM), Classification, GLCM (Gray Level Co-occurrence Matrix) Texture Features, deep semantic features

I. INTRODUCTION

Gastric cancer is one of the significant global deaths causing cancer diseases. After lung cancer, it is the second leading cause of mortality from cancer [1, 2]. The cause is carcinomas (Stomach cancer with an adenocarcinoma origin in mucus-producing cells). Among stomach cancers, this one is the most prevalent. Carcinoma gastric cancers make up the majority of malignancies that develop in the stomach, which make up more than 90% of tumours. The typical 5-year survival rate is less than 20%, which is a poor prognosis primarily due to late detection since the early stages are clinically quiet. Only a few nations, notably Japan, have established large early detection programs. The 5-year

survival rate can reach 90% if the tumour is found and treated before it invades the muscle layer of the stomach.[3]

Every year, more than 989,600 new diagnoses are made. 738,000 deaths are estimated to have occurred in 2008. 2012 saw an estimated 720 000 cases of stomach cancer fatalities [1]. There were 691 000 new instances of true gastric adenocarcinomas (non-cardia gastric cancers) in 2012, and 260 000 new cases of gastro-oesophageal junction adenocarcinomas (cardia gastric cancers), which are physically distinct. Despite significant improvements in our knowledge of epidemiology, pathology, molecular causes, and therapeutic options and tactics, as well as a drop in incidence and death, the burden of disease remains high [4]. Following lung and liver cancers, which, respectively, account for 23% and 28% of the total global burden of disability-adjusted life-years from cancer in men, gastric cancer accounts for 20% of the total [5]. Over 70% of new cases and deaths occur in developing countries. East Asia (Korea, Mongolia, Japan, and China) has the highest incidence rates, with yearly incidence rates ranging between 40 and 60 per 100,000 people. Contrary to the significantly lower rates recorded for the coastal and river valley regions, pockets of high risk are reported in Latin America's Andes Mountains, with rates between 20 and 30 per 100,000 [6]. Affluent populations in North America and Africa have lower rates (between 0.3 and 3 per 100,000). The rates of infection among African Americans are around twice as high as those among white Americans. In general, men have twice as many cases as women do. Different food habits, especially in European nations, and the incidence of *Helicobacter pylori* infection contribute to regional variations. The long-term low and steady prevalence of *Helicobacter pylori* infection in these nations likely explains this stalling of progress [7].

This study investigates to diagnose the early and advanced stages of gastric cancer using cutting edge Artificial intelligence-based techniques. Some of the researchers have used GLCM and deep semantic features for breast cancer diagnosis [8]. To improve diagnostic precision of early and advanced stages of gastric cancer carcinoma, we make also use of Gray-Level Co-occurrence Matrix (GLCM) texture features, deep semantic features and use machine learning classifier for the classification purpose. Our research hopes to

aid in the diagnosis of stomach cancer, resulting in more efficient treatment plans and higher patient survival rates.

II. MATERIALS AND METHODS

In this section, we elucidate the materials and methods used in this research study which includes data collection, data processing and our proposed approach for gastric cancer classification model.

A. Dataset Description

In our study, we used a publicly available gastric cancer carcinoma dataset. The dataset contains more than 1800 patch images with image size 512 X 512. The dataset contains images of early stages and advanced stages cancer. We chose 977 images of early and advanced stages.

B. Data Preprocessing

Data preparation is one of the key methods used in deep learning and machine learning. The degree to which the data has been pre-processed typically determines how accurate your model will be. In other words, the quality of your model is strongly correlated with the preprocessing of the data. We pre-processed the dataset for early and advanced stages. We used around 977 patch images of size 512 X 512. First, we convert the class labels into numerical labels representing 0 as early stage and 1 as advanced stage. The numerical

representation is required for the model to be used during the training and prediction phases. This is an integral step while training the model where class labels are non-numeric. Second, we extracted Gray level co-occurrence Matrix (GLCM). Third, we applied CNN based feature extractor on the data and extracted deep semantic features from the data. then, we combined these two kinds of features. The dataset was split into 80 training and 20% testing and finally, we trained the machine learning classifiers for early and advanced stages of gastric cancer carcinoma.

C. Proposed model

Features play a vital role in the performance of any machine learning or deep learning model. In this study, we aim to develop a model by combining the gray level co-occurrence matrix features and deep semantic features. The nature of both features is different. Mostly researchers use Convolutional Neural Networks (CNN) models as feature extractors, doing so the texture features are get less attention and are extracted from shallow layers though they are also very important for model training. Both features i.e., the semantic features which comes from CNN based extractors and the GLCM features which are basically the texture features representing the shapes, edges of images etc, have their own importance. The CNN features, which are often called deep semantic features, contain more spatial

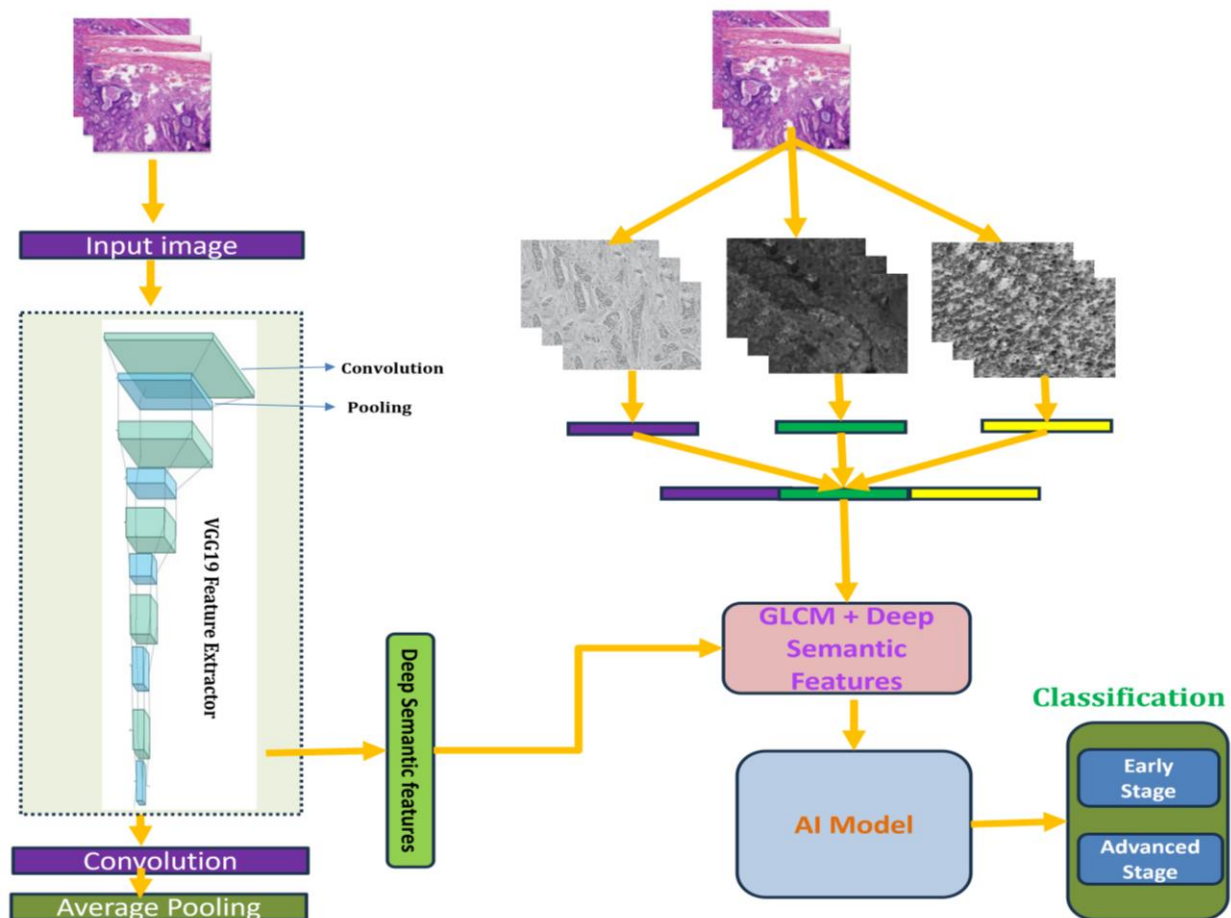


Figure 1. The architecture of proposed model.

information and gives more detailed information about the distribution of tissues. GLCM illustrates the texture of images. Herlick et al. [9] proposed GLCM to describe texture features. We calculated 15 different GLCM features including energy, contrast homogeneity, correlation, dissimilarity etc. For extracting deep semantic features, we used VGG19 and extracted 131087 features. We concatenated both the GLCM features and deep semantic features. Then we trained machine learning models with these combined features. The architecture of our proposed model has been shown in figure 1.

III. EXPERIMENTAL RESULTS AND DISCUSSION

This section explains about the experimental results about this research work. We extracted deep semantic features and GLCM features and concatenated both features and used three machine learning classifiers for the classification purpose of the different stages of gastric carcinoma. Our experimental results revealed that SVM outperformed other two models and demonstrated 96.93 % accuracy. Table 1 shows the performance metrics of all the applied models.

TABLE 1. PERFORMANCE MEASURES OF THE MODELS

Model	Accuracy	Precision	Recall	F1-score
SVM	96.93	96.01	96.96	96.05
DT	94.89	94.56	94.01	94.86
KNN	94.16	94.00	94.23	94.11

Note: SVM denotes support vector machine; DT denotes decision tree and KNN denotes K nearest neighbours.

There are several metrics for the evaluation of the model. One of them is confusion matrix. Figure 2 shows the confusion matrix of our model showing the distribution of all the predicted responses to their true classes. Out of 196 test samples, 190 test samples as shown on the diagonal of the confusion matrix are correctly predicted and 6 samples are misclassified by the model. It means the model accuracy is quite encouraging for classifying the stages of gastric cancer carcinoma. Likewise figure 3 shows the ROC curve.

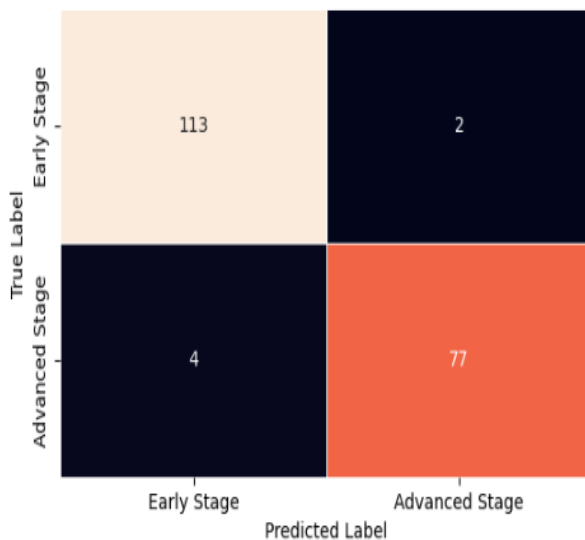


Figure 2. Confusion matrix of proposed model.

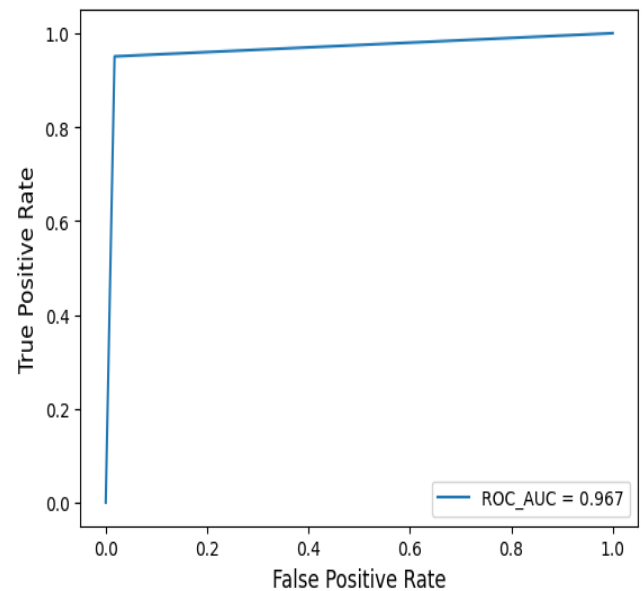


Figure 3. ROC curve of proposed model

Stomach cancer has been investigated by using different approaches over the years. For example Sakai et al.[10] proposed an automatic detection method using convolutional neural networks to help identify early stomach cancer in endoscopic images. Using two classes (cancer and normal) of image datasets with comprehensive texture information on lesions acquired from a limited number of annotated images, he performed transfer learning. The trained model's accuracy was 87.6%. Furthermore, he produced a heat map of unknown images that appeared to be a candidate zone of early stomach cancer. The accuracy of detection was 82.8%.

Sharma et al.[11] proposed a model where H&E stained histological whole slide images of gastric cancer are used to investigate deep learning techniques. In order to classify cancers based on immunohistochemical reaction and necrosis detection, a convolutional neural network is introduced. Handcrafted features are used to compare the deep learning methodology with conventional methods. The findings surpass conventional methods with an overall classification accuracy of 0.6990 for malignancy and 0.8144 for necrosis detection.

Li et al.[12] presented a study, a deep learning radiomics nomogram based on DECT which demonstrated high predictive value in identifying the location of lymph node metastases (LNM) in gastric cancer patients. This nomogram achieved a noteworthy accuracy of 0.77 (training set) and 0.76 (test set), outperforming both single-energy and clinical models. The best-performing model for the radiomics nomogram was chosen after a comparison of four models. The nomogram demonstrated strong predictive power for patient survival by including CT-reported lymph node status, radiomics signatures, and clinical data. This work enhanced the field of radiomics research on gastric cancer by offering a very precise method for prognosticating patients and predicting LNM.

Yang et al. [13] presented a study, based on publicly accessible gene expression data from TCGA-STAD, they

compared and assessed three machine learning models to predict metastasis of Lymph nodes. The features were chosen based on their associations with the LN status using the Pearson correlation coefficient (PCC) technique. The accuracy and F1 score were used to evaluate the model's performance. Using 26 specifically selected gene features, the Naive Bayesian model performed better, with an accuracy of 0.72 in the test set and 0.741 in the training set. In both the training and test sets, the F1 score was 0.597 and 0.652, respectively.

Mortezagholi et al. [14] presented a comparative analysis between the two groups of participants the healthy and the sick, they chose 405 samples. 11 traits and risk variables in all were looked at. To categorize the patients with stomach cancer, they employed four machine learning techniques: k-nearest-neighborhood (KNN), naïve Bayesian model, decision tree (DT), and support vector machine (SVM). The accuracy rates of the SVM, DT, naïve Bayesian model, and KNN algorithms were 90.08, 87.89, 87.60, and 87.60 percent, respectively, based on the outcomes of the evaluation of the four approaches. The results demonstrated that the KNN algorithm (87.17) had the lowest rate and the SVM (91.99) had the highest level of F-Score.

IV. CONCLUSIONS

In the past few years, gastric cancer remained one of the world's most difficult health problems. Artificial technologies have been developed for the prediction, diagnosis, and prognosis of this disease aiming to address this issue. In this study, the early and advanced stages of gastric cancer carcinoma was investigated by combining GLCM features and deep semantic features. For the classification of the different cancer stages, machine learning classifiers were used while training the models with combined features. We employed Support vector machine (SVM), Decision Tree (DT), and K-nearest neighbour (KNN) machine learning models for this purpose to determine the cancer stage outcomes of gastric cancer patients. SVM outperformed other two classifiers demonstrated higher results for identifying gastric cancer carcinoma stages. This approach would aid in the exact and accurate diagnosis of the gastric cancer in clinical applications. This research study is a part of our major research project. We aim to investigate more approaches and techniques and explore the diagnosis of gastric cancer carcinoma.

ACKNOWLEDGMENT

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977).

REFERENCES

- [1] Ferlay, J., et al., Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European journal of cancer*, 2013. 49(6): p. 1374-1403.
- [2] Jemal, A., et al., Global cancer statistics. *CA: a cancer journal for clinicians*, 2011. 61(2): p. 69-90.
- [3] Miyahara, R., et al., Prevalence and prognosis of gastric cancer detected by screening in a large Japanese population: data from a single institute over 30 years. *Journal of gastroenterology and hepatology*, 2007. 22(9): p. 1435-1442.
- [4] Colquhoun, A., et al., Global patterns of cardia and non-cardia gastric cancer incidence in 2012. *Gut*, 2015. 64(12): p. 1881-1888.
- [5] Soerjomataram, I., et al., Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. *The Lancet*, 2012. 380(9856): p. 1840-1850.
- [6] Correa, P., et al., Gastric cancer in Colombia. III. Natural history of precursor lesions. *Journal of the National Cancer Institute*, 1976. 57(5): p. 1027-1035.
- [7] Ferro, A., et al., Worldwide trends in gastric cancer mortality (1980–2011), with predictions to 2015, and incidence by subtype. *European journal of cancer*, 2014. 50(7): p. 1330-1344.
- [8] Hao, Yan, et al. "Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix." *Plos one* 17.5 (2022): e0267955.
- [9] Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 1973; 6, 610–621.
- [10] Sakai, Y., et al. Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018. IEEE.
- [11] Sharma, H., et al., Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Computerized Medical Imaging and Graphics*, 2017. 61: p. 2-13.
- [12] Li, J., et al., Dual-energy CT-based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer. *European Radiology*, 2020. 30: p. 2324-2333.
- [13] Yang, Y., et al., An immune-related gene panel for preoperative lymph node status evaluation in advanced gastric cancer. *BioMed Research International*, 2020. 2020.
- [14] Mortezagholi, A., et al., Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: Using data mining method. *Asian Pacific journal of cancer prevention: APJCP*, 2019. 20(9): p. 2607.



Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chungbuk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea majoring in Artificial Intelligence in

healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.



Samman Fatima received her bachelor's degree in software engineering from Fatima Jinnah women University Pakistan. Now she is pursuing her Master's degree from Inje University. Her research interests are medical imaging, artificial intelligence, Healthcare data, machine learning, deep learning.



Ali Hussain received the B.S. degree in computer science from Government College University Faisalabad (GCUF), Pakistan, in 2019. He is currently pursuing the master's degree with the Department of Digital Anti-Aging Healthcare, Inje University, South Korea. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, computer vision, reinforcement learning, and medical imaging.



Maisam Ali received his B.E degree in Electrical and communication Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision.



Muhammad Yaseen received his B.E degree in Electrical Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision and medical imaging.



Tagne Poupi Theodore Armand was born in Cameroon, received Msc in information System and networking at ICT University USA, Cameroon Campus in 2021. Currently, he is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. His research interest field includes image processing with a focus on medical image analysis, Deep Learning, Machine Learning and Metaverse.



Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in SoGang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research

interests include machine learning, deep learning, Computer vision.